NCS TIB 91-6

NATIONAL COMMUNICATIONS SYSTEM

TECHNICAL INFORMATION BULLETIN 91-6

SUBJECTIVE TESTS OF TELECONFERENCING CODECS

JANUARY 1991

OFFICE OF THE MANAGER
NATIONAL COMMUNICATIONS SYSTEM

WASHINGTON, D.C. 20305

Jan 1991 Final

Subjective Tests of Teleconferencing Codecs

C-DCA100-87-C-0078

Delta Information Systems, Inc.
Horsham Business Center
Bldg. 3
300 Welsh Road
Horsham, PA 19044

National Communications System
Office of Technology & Standards
701 S. Court House Road
Arlington, VA 22204-2198

NCS TIB 91-6

Approved for Public Release; distribution unlimited.

This report describes the results of subjective evaluation of teleconferencing codecs. Preliminary tests used a previously developed test tape simulating typical teleconferencing applications which resulted in a quality rating of each bit rate for the selected applications. Later tests emphasized motion performance using the more recently developed Video Codec Test Tapes Limited Motion and Full Motion. These tapes were taken to the location of each manufacturer and fed into the codec to be tested. The manufacturer then could certify that this codec was in proper operation condition. The codecs were connected back-to-back and processed output tapes obtained at 64, 128, 256, 384, 768, and 1536 (or 1544) Kbps, in each case covering the full operating range of the codec. The processed tapes were viewed and rated in terms of picture impairment by a group of impartial evaluators. The test scores obtained from all evaluators were analyzed and finally correlated with previously obtained initial objective test results.

Teleconferencing codecs
Video codecs

58

Unclassified Unclassified Unclassified Unlimited

Block

Block

Block

NCS TECHNICAL INFORMATION BULLETIN 91-6

SUBJECTIVE TESTS OF TELECONFERENCING CODECS

JANUARY 1991

PROJECT OFFICER

GARY M. REKSTAD
Electronics Engineer
Office of NCS Technology
and Standards

APPROVED FOR PUBLICATION:

Dennis Bodson

DENNIS BODSON
Assistant Manager
Office of NCS Technology
and Standards

FOREWORD

Among the responsibilities assigned to the Office of the Manager, National Communications System, is the management of the Federal Telecommunication Standards Program. Under this program, the NCS, with the assistance of the Federal Telecommunication Standards Committee identified, develops, and coordinates proposed Federal Standards which either contribute to the interoperability of functionally similar Federal telecommunication systems or to the achievement of a compatible and efficient interface between computer and telecommunication systems. In developing and coordinating these standards, a considerable amount of effort is expended in initiating and pursuing joint standards development efforts with appropriate technical committees of the International Organization for Standardization, and the International Telegraph and Telephone Consultative Committee of the International Telecommunication Union. This Technical Information Bulletin presents and overview of an effort which is contributing to the development of compatible Federal, national, and international standards in the area of teleconferencing. It has been prepared to inform interested Federal activities of the progress of these efforts. Any comments, inputs or statements of requirements which could assist in the advancement of this work are welcome and should be addressed to:

Office of the Manager
National Communications System
ATTN: NCS-TS
Washington, DC 20305-2010

SUBJECTIVE TESTS
OF TELECONFERENCING
CODECS

January, 1991

FINAL REPORT
DCA100-87-C-0078
TASK ORDER NUMBER 88-006

Submitted to:
NATIONAL COMMUNICATIONS SYSTEM
WASHINGTON, DC

DELTA INFORMATION SYSTEMS, INC.
300 Welsh Road, Ste. 120
Horsham, PA 19044-2273
TEL: (215) 657-5270 FAX: (215) 657-5273

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION AND SUMMARY

This document summarizes work performed by Delta Information Systems, Inc. (DIS) for the National Communications System (NCS), Office of Technology and Standards. This office is responsible for the management of the Federal Telecommunications Standards Program, which develops telecommunications standards, whose use is mandatory for all Federal departments and agencies. This study was performed under task order number 88-006 of contract number DCA100-87-C-0078.

This report describes the results of subjective evaluation of teleconferencing codecs. Preliminary tests used a previously developed test tape simulating typical teleconferencing applications which resulted in a quality rating of each bit rate for the selected applications. Later tests emphasized motion performance using the more recently developed Video Codec Test Tape, Part C: Limited Motion, and Part D: Full Motion. These tapes were taken to the location of each manufacturer and fed into the codec to be tested. The manufacturer thus could certify that his codec was in proper operating condition. DIS specified all required test equipment which in most cases could be made available by the manufacturer but was supplied by DIS if needed. The codecs were connected back-to-back and processed output tapes obtained at 64, 128, 256, 384, 768, and 1536 (or 1544) Kbps, in each case covering the full operating range of the codec. The processed tapes were viewed and rated in terms of picture impairment by a group of impartial evaluators. The test scores obtained from all evaluators were analyzed and finally correlated with previously obtained initial objective test results.

Sectic 2.0 of this report describes the available background material consisting of both previous NCS programs and related efforts performed for other parties. Section 3.0 establishes the various criteria to be considered in the evaluation of codecs. The variety of possible subjective test methods are discussed in Section 4.0. The contents of this section are primarily based on CCIR Recommendation 500-3 but stress variations due to the unique properties of teleconferencing codecs. Section 5.0 describes a series of preliminary tests with a tape which was specifically arranged to cover various user applications. Section 6.0 comprises the main portion of the test program. It discusses the test material and details of the test implementation. Close attention is given to the scheduling of the tests. The results and their subsequent analysis are discussed in detail. Finally, an initial check of correlation between subjective and objective test results is made. Section 7.0 summarizes the total program and gives a variety of suggested future efforts to produce a convenient integrated test program for digital video teleconferencing codecs.

2.0 BACKGROUND

2.1 Previous NCS Programs

Two previous programs directly provide material used for this program. They are Development of a Video Tape to Test Video Codecs Operating at 64 Kbps, NCS-TIB-89-2, and Development of a Video Tape to Correlate Subjective and Objective Testing of Teleconference Systems, NCS-TIB-90-7. The test tapes produced in these two programs provided the mainstay of the test material used in this program which fully proved their adequacy and usefulness.

Other related programs are: Standardization of End-to-End Performance for Full Motion Video Teleconferencing, NCS-TIB-90-2 and Analysis of Temporal Frequency Response as a Technique to Measure the Ability of a Teleconferencing System to Reproduce Motion, Contract No. DCA100-87-C-0078, Task Order No. 88-009, documented in a still to be published final report. These two tasks were primarily concerned with methods of objective testing but are nevertheless important because the ultimate goal of the various test programs is to develop a comprehensive approach to the testing of all digital video codecs.

2.2 INTELSAT Program

In May, 1986, Delta Information Systems performed a major digital video codec testing program for INTELSAT during the Seventh International Conference of Digital Communications by Satellite (IDCSC-7) in Munich, Germany. Six different codecs covering the range from 56 Kbps to 30 Mbps operating in NTSC and PAL standards and manufactured in three different countries were available for testing by over 150 conference participants from all over the world which included about 10% teleconferencing experts.

The stated purpose of the tests was to assess the usability of a codec operating at a given bit rate for several specific applications. This required a specially designed test tape and explicit instructions to the evaluators. The results proved to be highly gratifying and useful. The methodology used for these INTELSAT tests is directly applicable to some portions of the present program.

3.0 EVALUATION CRITERIA

3.1 Codec Characteristics

Every digital video codec using bandwidth compression inherently produces an impaired picture. The type and severity of these impairments depends on the codec algorithm and the bit rate at which it operates. This will remain true into the foreseeable future but the CCITT H.261 Recommendation just being introduced on a worldwide basis will force a degree of uniformity on the industry. The picture impairments manifest themselves as various artifacts which can often be distinguished and identified by a video codec expert but not by an average teleconferencing user who judges the picture on its overall impression. Codec design allows some latitude for trade-offs between artifacts. The evaluation scores given by non-expert observers show the combined effects of the various codec design features.

3.2 Test Picture Features

All test sequences used for codec performance evaluation have been designed to determine the codec's response to certain typical features in various transmitted pictures. Indeed, the most recent series of test tapes has been divided into four parts each of which emphasizes requirements for resolution, motion rendition and combinations of both. Depending on the purpose of a specific test, only selected parts of the test tape are used. Each part contains a sufficient variety of partly rather similar sequences so that a valid average can be obtained and the need to evaluate each sequence more than once can be eliminated. It is often suggested to check the consistency of the ratings of each evaluator by presenting each sequence twice but this adds to boredom and fatigue and is better avoided. Different test sequences "exercise" the performance parameters of any codec and yield a general average quality rating.

3.3 Application Groups

Wherever digital video transmission is implemented, it is to serve a distinct purpose for a specific application. This may impose widely differing performance requirements, mainly in the areas of picture resolution and motion rendition. Numerous attempts have been made by technical committees of organizations such as CCITT and the American National Standards Institute (ANSI) to categorize various applications and their required levels of service but so far no consensus has been achieved to bring them down to a sufficiently universal yet practically manageable list.

One series of tests covered in this report has been performed in terms of four selected application groups, others in terms of general picture qualitv. Once a firm categorization of application groups has been made by the standards organizations and accepted by the communications industry, it will be possible to examine each sequence on the test tapes and assign it as applicable to one or several application groups. By this means it will be possible to process the test data obtained for this report in different groupings and achieve ratings for various applications.

4.0 TEST METHODOLOGY

4.1 Quality Assessment Scales

The fundamentals of a standard method of subjective evaluation of the quality of television pictures has been laid down in CCIR Recommendation 500-3 and its associated Report 405-5. It is based on a group of preferably non-expert observers viewing and rating the appearance of the codec output picture. In the case under discussion here, only a person directly familiar with video codecs would be considered an expert.

In most cases, the single stimulus method is used, meaning that only one picture is presented at one time. This picture can be rated based on either quality or impairment as shown here.

Five-grade scale

Quality	Impairment
5 Excellent	5 Imperceptible
4 Good	4 Perceptible, but not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

Six-grade scales have been used but are now obsolete. In case two codecs are to be compared, the double stimulus method with two monitors displaying synchronized pictures is applicable using the comparison scale shown below.

+3	Much better
+2	Better
+1	Slightly better
0	The same
-1	Slightly worse
-2	Worse
-3	Much worse

A double stimulus display with the original picture on one monitor can also be used with the quality or impairment scales but this is not very common.

Obviously the assessment of quality or impairment is mainly and should be up to the judgement of each observer without any outside influence. However, in order to achieve some consistency in the ratings, some guidance is very desirable. It is logical and easy to always show the original as reference but presenting a sample of a degraded picture can be much more of a problem. It can be done with reasonable confidence with a picture of or near broadcast quality where the expected impairments are mainly loss of resolution and added noise. These impairments can easily be generated and added to the picture in a measured amount to produce a "worst case" picture to serve as a so-called "anchor". However, a codec output picture can have too many severe and radically different impairments that it would be impossible to artificially produce a standard "poor" or "bad" picture. Therefore, the rating depends exclusively on the individual judgement of each observer. It is possible that somebody may, for instance, give a picture the lowest rating at an early test and later encounter an even worse picture. The effect of this potential problem can be minimized by judicious scheduling of the same tests at different times with different observer teams, in addition to the benefit of averaging over a sufficient number of test sequences and observers.

4.2 Application to Codec Testing

CCIR Recommendation 500-3 and other closely related more specific test procedures were all developed for use with fairly high quality pictures with limited impairments. They are readily applicable to analog broadcast TV and also high quality digital systems designed for transmission over DS-3 or broadband ISDN circuits. However, the teleconferencing codecs tested in this program are limited to transmission over basic or at the most primary rate ISDN channels which inherently produces much more severely impaired pictures and imposes the need for some modifications of the test and evaluation procedures.

Quoting CCIR Rec. 500-3, "When using the quality or impairment scale, the range of impairments should be chosen, wherever practicable, so that all grades are used by the majority of observers; a grand mean score (averaged over all judgements made in the experiment) close to 3 should be aimed at, to standardize results." This objective cannot be fulfilled with teleconferencing codecs at any one bit rate because the resulting picture quality varies over too wide a range. However, the grand mean score over the whole range of operating bit rates of one codec model may at least approach the above objective.

CCIR Rec. 500-3 also contains an elaborate list of preferred viewing conditions which are recommended but not compulsory. In the case of evaluating high quality pictures where small differences become important, these conditions should be adhered to quite closely. The wide range of impairments of teleconferencing codecs, however, is easily discernible under most viewing conditions. Nevertheless, the Rec. 500-3 preferred conditions should be complied with as much as practical.

4.3 Evaluation Techniques

Though several evaluation techniques employing different levels of sophistication have been proposed and used, the most frequent and straightforward method is to compute the mean scores for each test sequence and then the mean of these values to arrive at an overall score for each codec at each operating bit rate. The standard deviations of these scores and the mean values for each evaluator are not used directly but provide summarized information about opinion distributions and alerts to major discrepancies which may call for special actions such as elimination of some scores, test sequences, and possibly evaluators. It is also feasible to re-arrange the scores to single out ratings for selected types of test sequences and specific application groups.

Unless specific instructions to the contrary were given to the evaluators (which is extremely unlikely), each score reflects the integrated overall impression of each evaluator. Different impairments generally produce different levels of annoyance in each evaluator resulting in a variety of scores to be averaged to achieve a final rating. The ultimate objective of codec evaluation is to develop completely objective measuring techniques which are easier to use and more accurate than subjective testing. Development of such techniques is still in its early stages, but it has already become clear that generally different methods have to be used to measure different artifacts such as smear, blocking, jerkiness, etc. At present, it is not known how such measurement results can be combined to achieve an integrated objective score which can simulate the results of subjective evaluation. Even though attempts are being made to correlate subjective and objective test scores, the results are only preliminary and will leave much room for improvement.

5.0 USER APPLICATION GROUP TESTS

5.1 Setup

A series of subjective tests was performed with a test tape edited to arrange conventional scenes with a large variety of contents into groups typical of a number of user applications. These groups are considere'J representative but by no means unique, many other arrangements are possible.

The test tape consists of four segments, one for each application to be evaluated, each divided into two or three parts. The approximate contents of each segment and part are as follows:

Segment 1: Face-to-face Video Teleconference (between offices)
Part 1: Single person, head and shoulders
Part 2: Graphics with motion, zoom and pointing

Segment 2: Group Video Teleconference (between conference rooms)
Part 1: Groups of 3 and 6 persons
Part 2: Groups divided into single and pairs of persons

Segment 3: Tele-Education (instructor to many distant classrooms)
Part 1: Explanation of printed circuit board
Part 2: Drawing on flip chart pad
Part 3: Animated computer graphics, illustration, drawing

Segment 4: Briefing (company executive to branch offices)
Part 1: Animated computer graphics, flow chart
Part 2: Person explaining view graphs and map

The tests were performed during a meeting of the T1Q1.5 Subworking Group on Video Teleconferencing/Video Telephone in Baltimore, MD on November 8, 1988. They consisted of the screening of the above described ¾" video tape which had been processed through three video codecs operating at different bit rates. Three monitors were provided throughout the meeting room for viewing by members of the Subworking Group who agreed to participate in the tests.

Table 5 - 1 shows the questionnaire given to the evaluators for each test.

TEST AND EVALUATION OF VIDEO CODECS
WHEN USED FOR TELECONFERENCING

Evaluator Name: ____

YES NO

1. I consider myself a technical expert in digital video teleconferencing. □ □

2. I have used video telecconferencing in the past. □ □

3. If the answer to No. 2 is yes, has the experience been favorable? □ □

RATINGS

Segment No.	Application Category	Part No.	Excel.	Good	Fair	Poor	Bad
1	Face-to-Face Video Telephone	1					
		2					
2	Group Teleconference	1					
		2					
3	Tele-education	1					
		2					
		3					
4	Briefing	1					
		2					

Test No.________

TABLE 5-1

EVALUATION QUESTIONNAIRE

Only 3 evaluators considered themselves technical experts but most of them had used video teleconferencing and regarded the experience as favorable. The quality rating scale of CCIR Rec. 500-3 was used. The recommended viewing conditions such as room illumination, monitor brightness and distance were implemented as closely as the meeting room environment would permit but were not ideal. Time constraints made it necessary to make all tests in the shortest possible time with minimal pauses which resulted in a total almost continuous test time of about 80 minutes which is longer than recommended. The only noticeable effect of these deviations from ideal conditions was the tendency of some evaluators to give higher ratings in later tests.

Six ¾" tapes were presented to the evaluators, consisting of the above described test tape after processing at six data rates through three different type codecs. The data rates were 64, 128, 256, 384, 768, and 1544 Kbps and were presented in a random order known only to the test director. As recommended in CCIR Rec. 500-3, most evaluators were not sufficiently expert in video codecs to identify the manufacturers which enhances the validity of the results. A short taped aural instruction to the evaluators stating the salient features of each represented application preceded the presentation of each segment.

5.2 Results

The individual test scores and the results of their numerical evaluation are given on Tables 5 - 2 to 5 - 7, one for each data rate under test. The scores are first averaged for each evaluator and part of the test tape. Subsequently, averages are computed for each application (segment) and the total for each data rate. Finally, special averages are computed for three predominant tape contents independent of applications, namely Persons (Parts 1-1, 2-1, 2-2, 3-2), Graphics (Parts 1-2, 3-3, 4-1), and combined Persons and Graphics (Parts 3-1, 4-2). All these results are shown in curves on Figures 5.1 to 5.8 in terms of score vs. bit rate.

The test scores show a typical feature of subjective tests, namely wide variations between evaluators. To check for erratic scoring by one evaluator or inconsistent results caused by one test part, standard deviations were computed for each line and column on Tables 5 - 2 to 5 - 7. All deviations were low enough (almost all below one) to confirm the validity of the results.

EVALUATOR NO.	SEGMENT	1		2		3			4		
	PART	1	2	1	2	1	2	3	1	2	AVERAGE
1		3	2	3	2	1	3	2	2	1	2.1
2		2	2	2	1.5	1	2	2	2	1	1.7
3		2	2	2	2	1	3	2	1	2	1.9
4		3	2	2	2	2	2	2	2	2	2.1
5		2	1	2	2	1	2	1	2	1	1.6
6		3	1	2	2	3	3	3	2	1	2.2
7		3	2	3	2	2	3	3	2	2	2.4
8		3	2	2	1	1	2	2	2	1	1.8
9		3	3	3	3	3	4	3	2	2	2.9
10		4	3	2	2	1	3	2	2	1	2.2
11		3	2	2	3	2	3	2	3	1	2.3
12		2	2	2	2	2	2	3	2	2	2.1
13		4	3	3	3	1	3	2	2	1	2.4
14		2	2	2	2	2	3	2	2	2	2.1
15		2	1	2	2	1	3	2	1	1	1.7
AVERAGES	PART	2.7	2.0	2.3	2.1	1.6	2.7	2.2	1.9	1.4	
	SEGMENT	2.3		2.2		2.2			1.7		
	TOTAL	2.1									
	PERSONS	2.4									
	GRAPHICS	2.0									
	PERSONS & GRAPHICS	1.5									

TABLE 5-2
TEST RESULTS - 64 KBPS

EVALUATOR NO.	SEGMENT	1		2		3			4		
	PART	1	2	1	2	1	2	3	1	2	AVERAGE
1		1	1	1	1	1	3	2	2	-	1.5
2		3	3	3	2.5	3	3	3	3	3	2.9
3		2	3	3	2	2	3	3	2	3	2.6
4		3	2	2	3	2	2	2	2	2	2.2
5		4	1	2	3	1	2	1	1	1	1.8
6		-	-	-	4	3	3	3	3	3	3.2
7		3	3	3	2	2	3	4	3	3	2.9
8		3	2	2	1	2	2	2	2	1	1.9
9		3	3	3	3	4	4	3	3	3	3.2
10		3	2	2	3	2	3	2	2	1	2.2
11		3	2	2	3	2	3	2	3	1	2.3
12		3	2	2	2	2	2	2	2	2	2.1
13		5	4	4	4	3	4	3	3	1	3.4
14		3	2	2	3	2	3	3	2	2	2.4
15		2	2	2	2	3	2	2	2	2	2.1
AVERAGES	PART	2.9	2.3	2.4	2.6	2.2	2.9	2.5	2.3	2.0	
	SEGMENT	2.6		2.5		2.6			2.2		
	TOTAL	2.4									
	PERSONS	2.7									
	GRAPHICS	2.4									
	PERSONS & GRAPHICS	2.1									

TABLE 5-3
TEST RESULTS - 128 KBPS

EVALUATOR NO.	SEGMENT	1		2		3			4		
	PART	1	2	1	2	1	2	3	1	2	AVERAGE
1		1	2	2	2	1	1	2	2	2	1.7
2		4	3	3.5	3.5	4	5	4	2.5	3	3.6
3		3	3	3	3	2	4	3	3	3	3.0
4		2	2	2	2	1	2	3	2	2	2.0
5		2	2	3	2	3	4	3	2	2	2.6
6		3	2	3	3	3	3	4	2	2	2.8
7		3	2	2	2	1	3	3	3	3	2.4
8		4	3	1	1	2	2	3	3	1	2.2
9		4	3	3	4	4	4	4	4	3	3.7
10		4	2	2	3	2	3	2	3	1	2.4
11		3	2	2	3	2	3	2	3	1	2.3
12		3	2	2	2	3	3	3	3	2	2.6
13		3	2	3	3	3	3	3	3	1	2.7
14		3	2	2	3	2	4	3	3	3	2.8
15		3	2	3	3	2	3	3	3	2	2.7
	PART	3.0	2.3	2.4	2.6	2.3	3.1	3.0	2.8	2.1	
	SEGMENT	2.6		2.5		2.7			2.5		
AVERAGES	TOTAL	2.6									
	PERSONS	2.8									
	GRAPHICS	2.7									
	PERSONS & GRAPHICS	2.2									

TABLE 5-4
TEST RESULTS - 256 KBPS

EVALUATOR NO.	SEGMENT	1		2		3			4		
	PART	1	2	1	2	1	2	3	1	2	AVERAGE
1		2	2	3	3	2	3	2	3	1	2.3
2		3	3	4	4	3	4	4	3	3	3.4
3		3	4	4	3	3	4	3	2	2	3.1
4		3	3	3	3	2	3	3	3	2	2.8
5		2	2	3	3	2	3	3	2	2	2.4
6		4	4	3	4	3	3	4	3	3	3.4
7		3	2	3	3	3	4	3	3	3	3.0
8		3	2	3	2	3	3	3	3	2	2.7
9		4	4	4	4	4	4	4	4	4	4.0
10		4	3	3	3	2	4	3	3	2	3.0
11		3	3	2	3	3	4	4	4	2	3.1
12		3	3	4	3	3	3	3	4	3	3.2
13		4	3	4	3	2	4	3	3	1	3.0
14		3	3	3	3	2	3	3	3	2	2.8
15		2	2	2	2	2	3	2	2	2	2.1
	PART	3.1	2.9	3.2	3.1	2.6	3.5	3.1	3.0	2.3	
	SEGMENT	3.0		3.1		3.1			2.7		
AVERAGES	TOTAL	3.0									
	PERSONS	3.2									
	GRAPHICS	3.0									
	PERSONS & GRAPHICS	2.5									

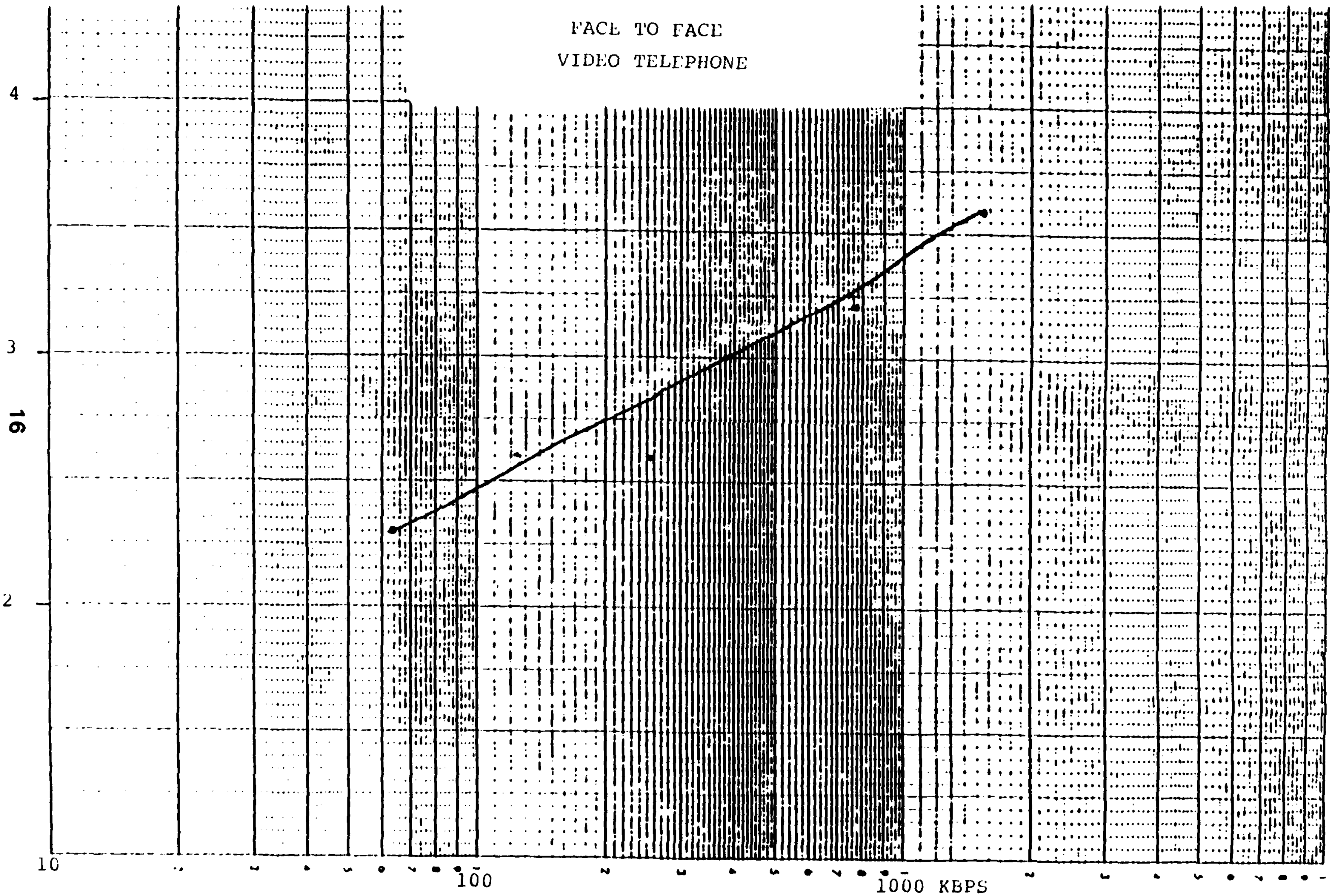
TABLE 5-5
TEST RESULTS - 384 KBPS

EVALUATOR NO.	SEGMENT	1		2		3			4		
	PART	1	2	1	2	1	2	3	1	2	AVERAGE
1		-	2	3	3	2	4	4	3	1	2.8
2		4	4	5	4	2.5	3	3.5	3	3.5	3.6
3		4	3	4	4	2	4	4	3	2	3.3
4		4	4	4	3	3	4	3	4	3	3.6
5		3	3	3	3	2	3	3	3	2	2.8
6		4	2	3	3	3	4	3	4	1	3.0
7		3	2	4	3	3	4	3	4	3	3.2
8		4	2	3	2	2	4	3	3	2	2.8
9		4	4	4	4	3	4	4	4	3	3.8
10		4	3	4	3	2	4	3	3	1	3.0
11		4	3	3	4	3	4	3	3	1	3.1
12		3	2	3	3	2	3	3	2	3	2.7
13		4	3	3	3	2	4	3	3	1	2.9
14		3	2	3	3	2	3	3	3	3	2.8
15		3	3	3	2	2	3	2	3	2	2.6
AVERAGES	PART	3.6	2.8	3.5	3.1	2.4	3.7	3.2	3.2	2.1	
	SEGMENT	3.2		3.3		3.1			2.7		
	TOTAL	3.1									
	PERSONS	3.5									
	GRAPHICS	3.1									
	PERSONS & GRAPHICS	2.3									

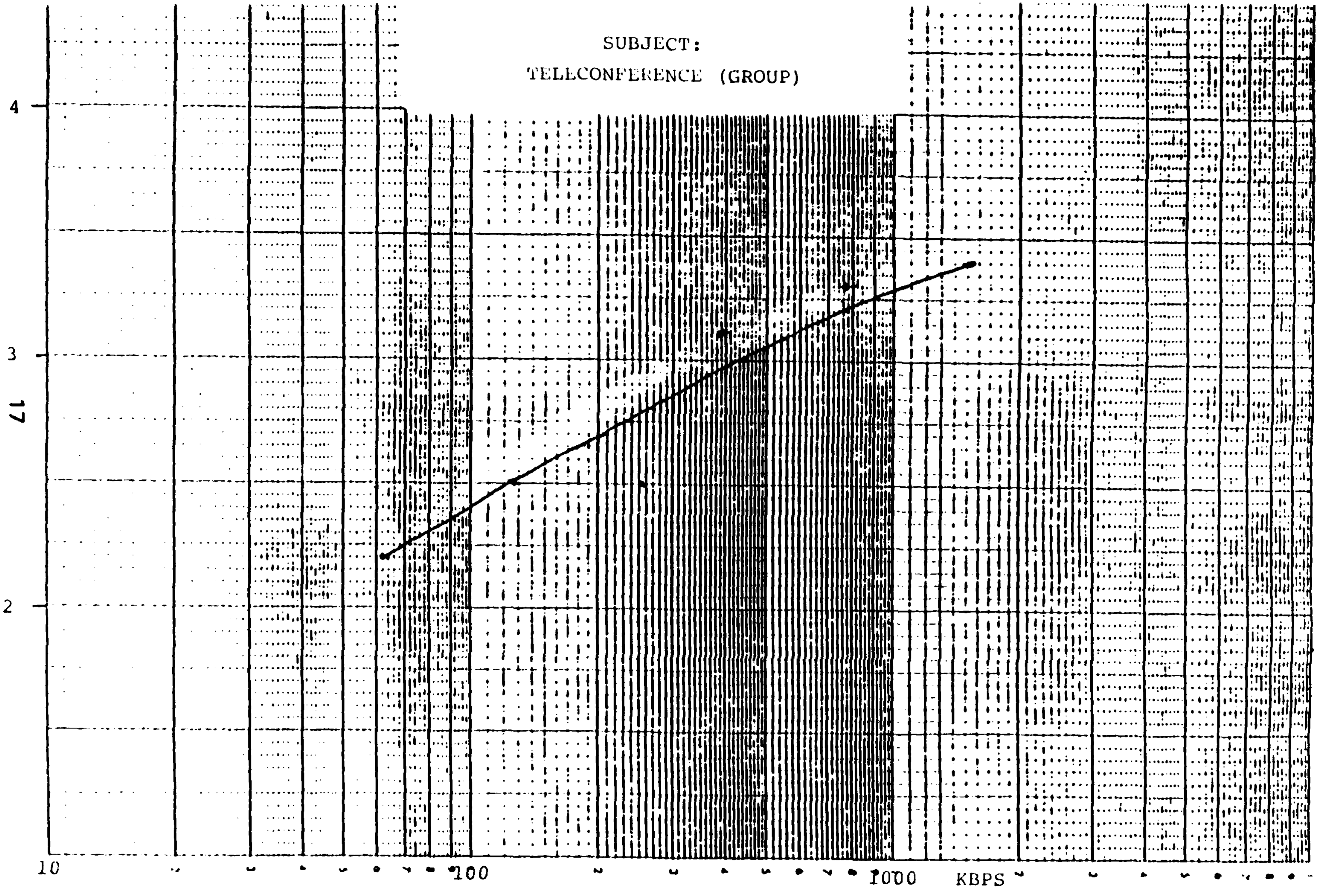
TABLE 5-6
TEST RESULTS - 768 KBPS

EVALUATOR NO.	SEGMENT	1		2		3			4		
	PART	1	2	1	2	1	2	3	1	2	AVERAGE
1		2	2	2	2	2	3	3	3	2	2.3
2		4	4	5	5	5	5	5	5	4	4.7
3		5	5	5	4	5	5	5	5	4	4.8
4		4	4	3	3	2	4	3	4	3	3.3
5		4	4	3	3	4	4	4	4	3	3.7
6		4	4	4	3	3	4	4	3	3	3.6
7		4	3	4	3	3	4	4	4	3	3.6
8		4	4	2	1	3	3	4	3	2	2.9
9		4	3	4	4	4	4	4	4	4	3.9
10		4	3	3	3	2	4	4	3	1	3.0
11		4	3	3	4	3	4	4	4	2	3.4
12		4	3	4	4	3	4	4	3	3	3.6
13		4	3	4	3	3	4	4	4	1	3.3
14		4	3	4	4	3	4	4	4	3	3.7
15		4	3	3	3	2	3	3	3	2	2.9
AVERAGES	PART	3.9	3.4	3.5	3.3	3.1	3.9	3.9	3.7	2.7	
	SEGMENT	3.6		3.4		3.6			3.2		
	TOTAL	3.5									
	PERSONS	3.6									
	GRAPHICS	3.7									
	PERSONS & GRAPHICS	2.9									

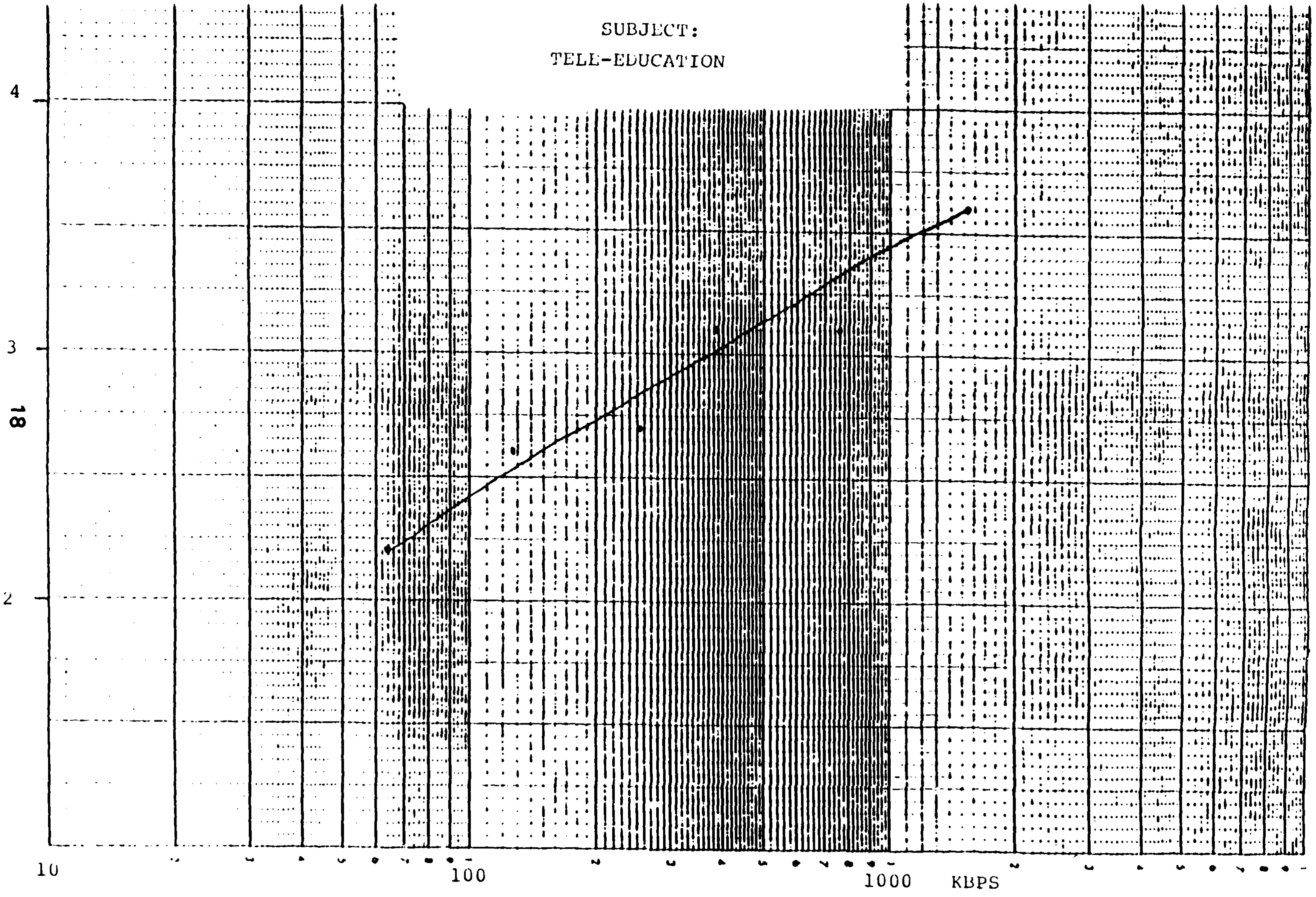
TABLE 5-7
TEST RESULTS - 1544 KBPS



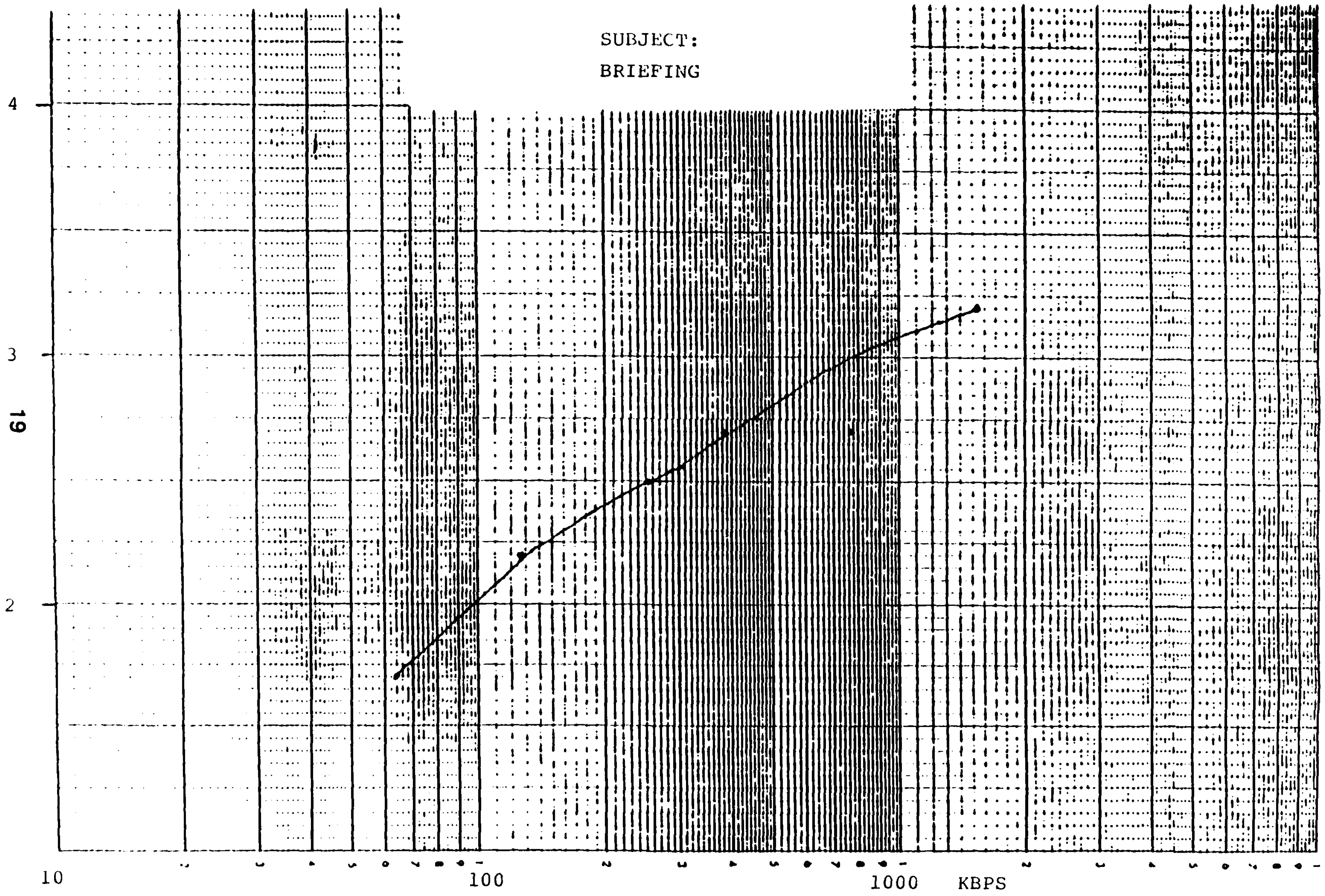
CODEC SUBJECTIVE RATING
FIGURE 5.1



CODEC SUBJECTIVE RATING
FIGURE 5.2

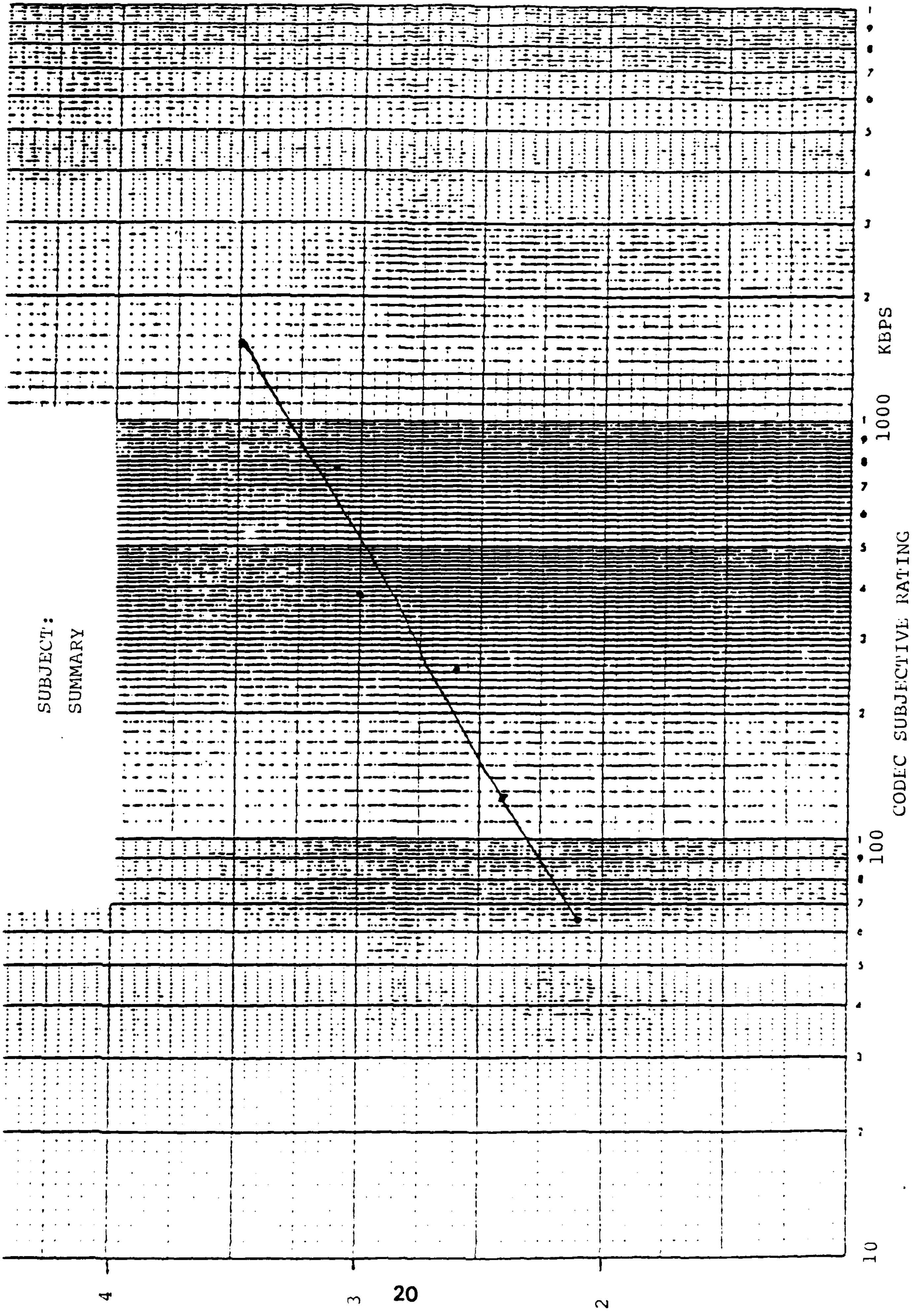


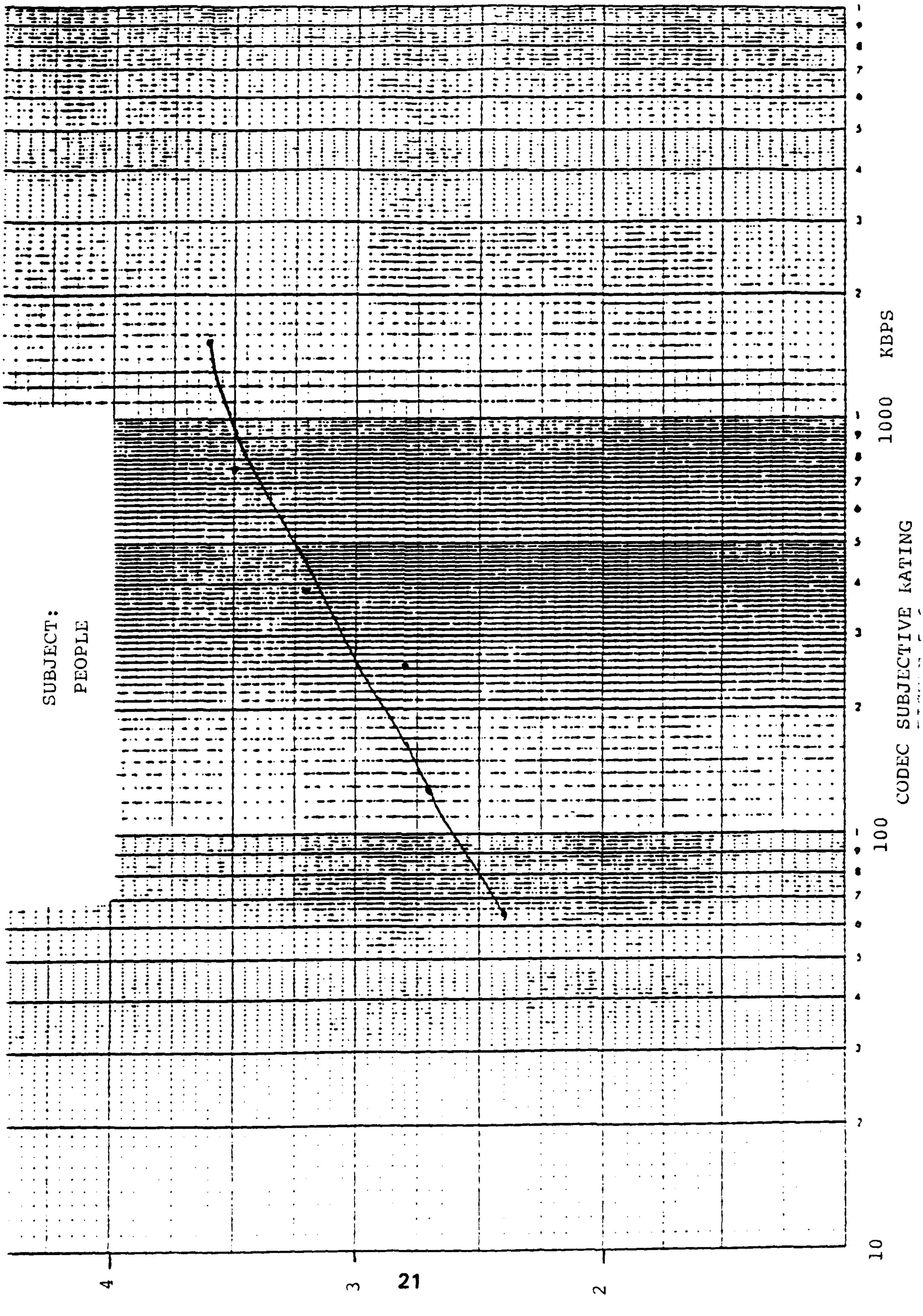
CODEC SUBJECTIVE RATING
FIGURE 5.[illegible]



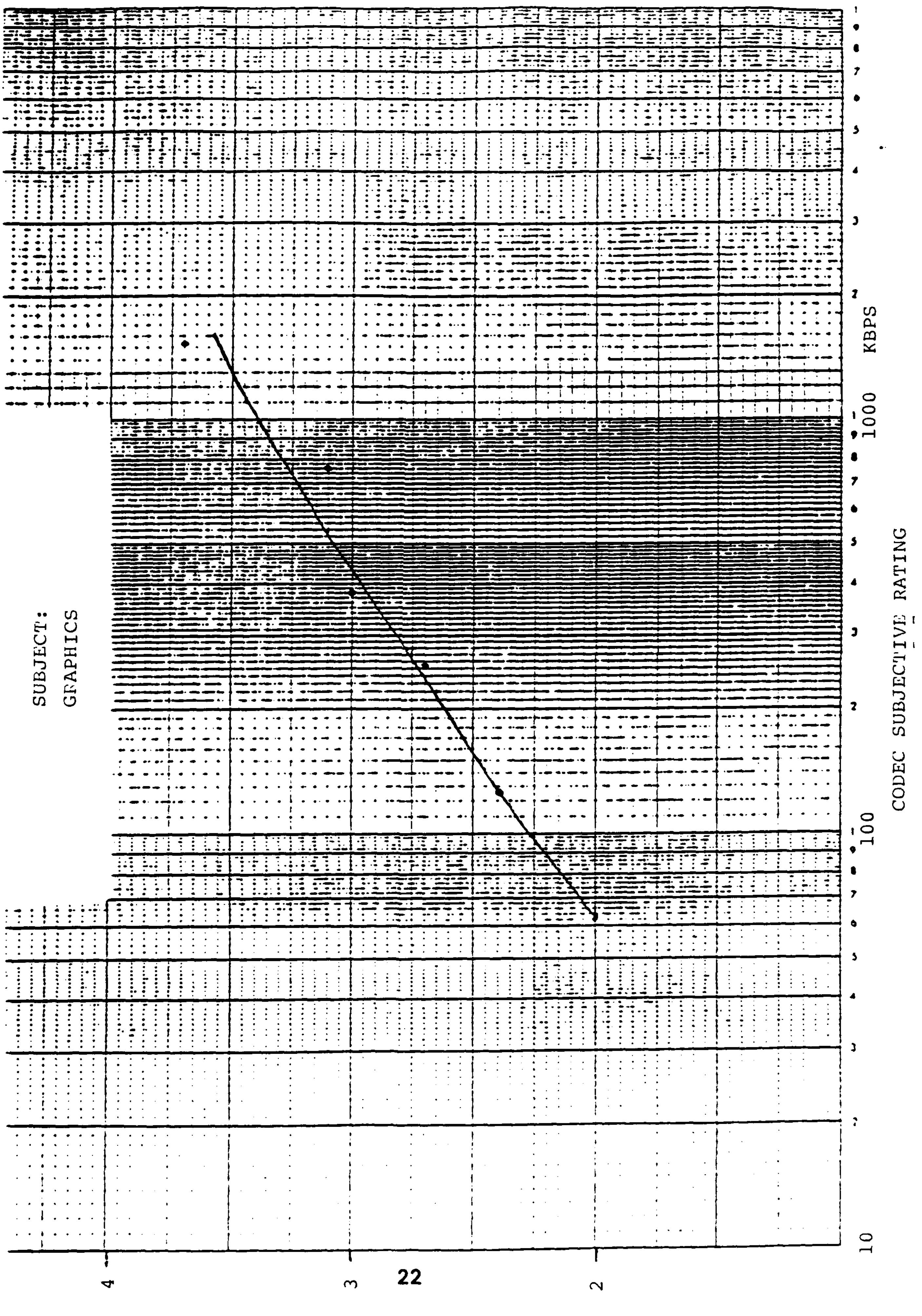
CODEC SUBJECTIVE RATING

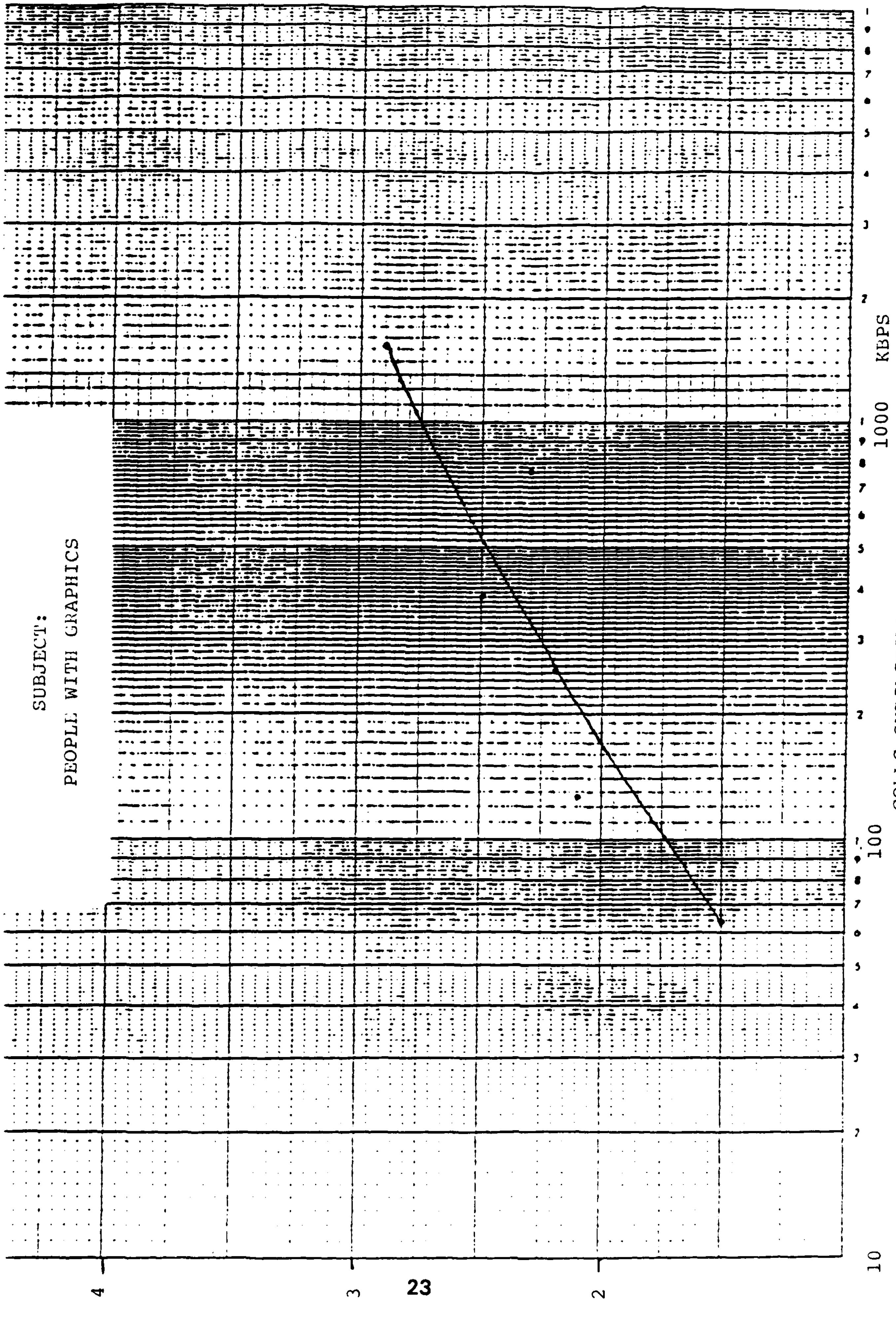
FIGURE 5.4


SUBJECT:
SUMMARY
CODEC SUBJECTIVE RATING
KBPS
10
100
1000
4
3
2



SUBJECT:
PEOPLE
CODEC SUBJECTIVE RATING
KBPS
10
100
1000
4
3
2


SUBJECT:
GRAPHICS
CODEC SUBJECTIVE RATING
KBPS
10
100
1000
4
3
2



CODEC SUBJECTIVE RATING

The overall range of all average scores extends from 1.5 to 3.5. A rating below 2 is assumed definitely unusable, between 2 and 2.5 marginal but any score above 2.5 should be considered usable for most applications. Based on these assumptions, a codec operating at 64 Kbps is marginally usable only for showing people in a teleconference environment where motion is limited and rendition of fine details is not required.

Rates of 128 Kbps and above appear to be usable for most applications. 384 Kbps give on the average fair results while the top rate of 1544 Kbps ranks between fair and good. This is the best that can be expected of any codec since some degradation is inevitable.

Pictures of people generally get the highest rating. Graphics become somewhat difficult when motion of the material is involved. The most difficult scenes are the ones showing people together with high detail graphics including camera panning and zooming, as exemplified by Parts 3-1 and 4-2. The same factors cause the briefing segment to be rated much lower than the other three applications which show only small differences. All curves confirm that the picture quality rating increases at higher bit rates. There are a few minor exceptions but they are no more than what must be expected considering the inherent uncertainties of subjective testing.

6.0 PICTURE QUALITY TESTS

6.1 Test Material

Previous experiments have shown clearly that the prime factor determining the picture quality of a digital teleconferencing video codec is its capability to reproduce motion. Conventional still picture parameters, primarily resolution, are definitely important, but they can be easily measured objectively by conventional methods. Motion performance, however, up to now can be evaluated only by subjective tests since objective methods are still in the development stage.

The series of tests described herein was implemented using the test tapes developed as part of the programs covered in NCS-TIB-89-2 and NCS-TIB-90-7 entitled Video Codec Test Tape, Part C: Limited Motion, and Part D: Full Motion. These tapes had previously been processed through three different model codecs. Two of them, designated L and P, operated in the low bit rate range at 64, 128, 256 and 384 Kbps, the third, designated H, in the high bit rate range at 384, 768 and 1536 Kbps. In view of the emphasis on motion performance, only Part C for codecs L and P, and Part D for codec H were used for this evaluation.

6.2 Implementation

The logical choice of test methodology is the single stimulus quality or impairment assessment scale. Either of these have been used most frequently for previous subjective picture quality tests. Even though such tests were generally used for the evaluation of high performance broadcast TV systems, essentially the same methods are applicable to digital teleconferencing codecs.

A method was agreed to by the European Broadcast Union (EBU) mainly for the assessment of high quality digital television pictures. This very popular method was also used in elaborate tests initiated by ANSI Committee T1Y1.1 to evaluate various algorithms for the transmission of digital TV at DS3 rates. The only significant difference in the procedure used for the tests described herein is that the impairment scale is more descriptive than the quality scale for rating the lower quality pictures produced by teleconferencing codecs.

Tapes C and D consist of respectively 16 and 18 sequences with durations ranging from 12 to 80 seconds, with most sequences lasting between 20 and 30 seconds. This results in a realistic crossection of the many types of scenes that n.ay occur in a teleconference or videophone application. The tape contains no audio, only a short "live" aural introduction was given at the start of the tests. Each sequence is first presented in its original form as reference, followed after a 3 second interval of medium grey (50 IRE) background by the processed sequence. Immediately following is the 10 second scoring interval which is visually identified with the sequence number to be scored. After a very short grey interval, the next reference sequence is presented. This timing arrangement including the tape recorder functions is graphically shown on Figure 6.1.

The tests were performed in a windowless room with light beige wells and easily controlled lighting. Five chairs for evaluators were provided in two rows, with the two chairs in front being about 4H and the three chairs in the rear about 6H distance from the monitor screen where H is the displayed picture height. This arrangement allowed all observers an unobstructed view and complied with CCIR REC 500-3. Light levels were kept close to the recommended values. A sketch of the room layout is shown on Figure 6.2.

Nine evaluators were available for the performance of the tests. Five were male and four female. All had professional training, partly with technical

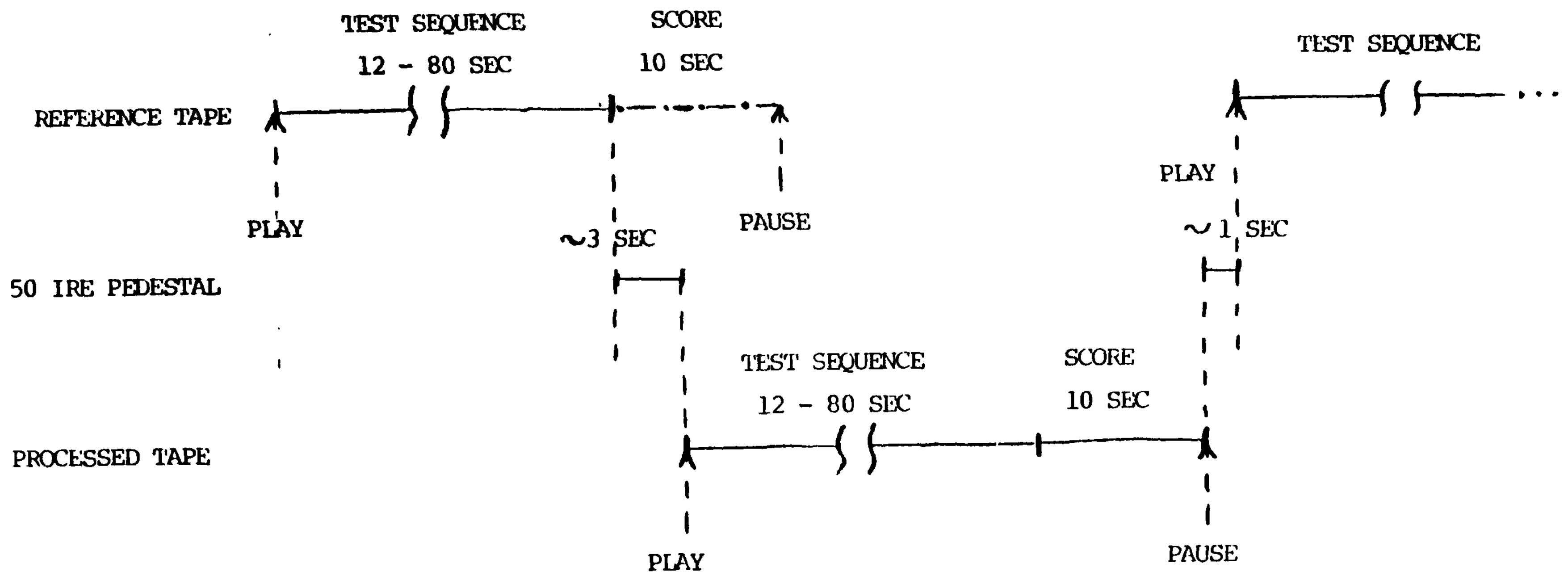


FIGURE 6.1
TEST TIMING

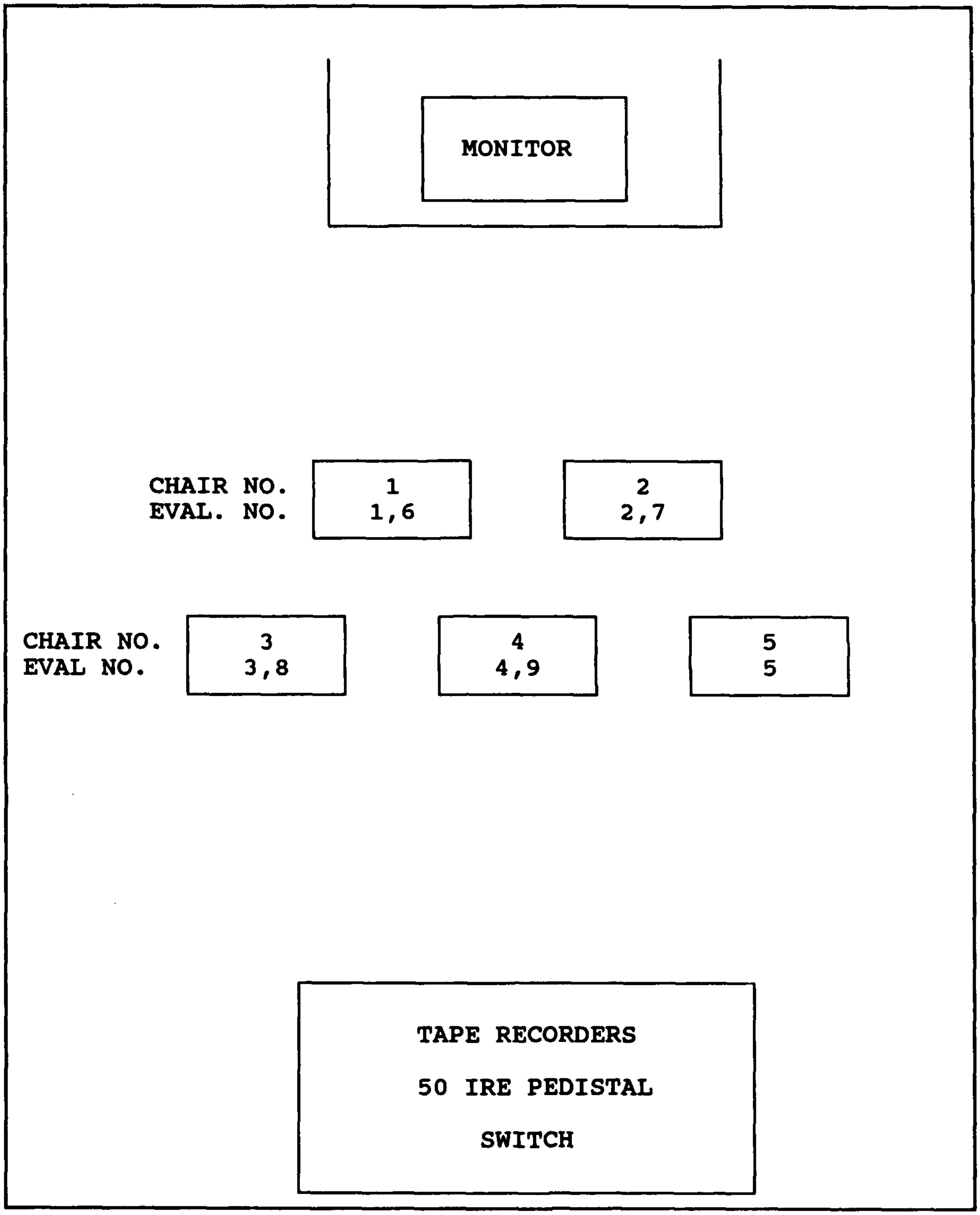


FIGURE 6.2
TEST FACILITY ROOM LAYOUT

background and some TV experience but none were experts in teleconferencing video codecs. Two groups, A and B, of five and four evaluators were formed, both containing male and female participants, identified by numbers which also indicated their seating locations. Chair #5 was not occupied in group B.

The score sheet used for each test is shown on Table 6-1. It identifies evaluator and test by number and contains a guide for impairment grading and a line for each test sequence. Previous experience has shown that many evaluators prefer a somewhat flexible scale, therefore allowance is made for scoring between the official five grades. In some cases, an essentially continuous scale with 0.1 point divisions has been suggested but that much detail is unnecessary and may actually lead to confusion.

6.3 Scheduling

Proper scheduling of a test series like the one to be performed here is very important to ensure both efficiency and fairness. The task of performance grading is stressful, so enough rest between sessions is needed to minimize fatigue. Even so, the reactions of an evaluator often differ between beginning and end of a session. External influences may produce day-to-day variations. Consecutive tests by the same evaluator should be dissimilar to avoid any possible interaction.

The running times of the tapes including titles and scoring intervals are about 12 minutes for Tape C and 16 minutes for Tape D. Including the reference runs of the unprocessed tape and the additional short intervals, the total running times reach about 23 minutes for Tape C and 29 minutes for Tape D which is well within the suggested limits of CCIR REC 500-3. An interval of between 10 and 15 minutes between sessions must be added for logistics purposes.

Since the tests were alternated between two groups of evaluators and 11 tapes were available, a total of 22 tests had to be performed. The intervals between tests by the same group were shorter than suggested in CCIR REC 500-3 but experience has shown that this caused no adverse effect. The resulting schedule as implemented is shown on Table 6-2. This schedule was known only to the test director, none of the evaluators had any knowledge of codec types or bit rates. The processed tapes were identified by numbers only. This schedule resulted in efficient use of facilities and personnel and satisfied all the above stated requirements.

CODEC EVALUATION FORM

EVALUATOR NO. __________ TEST NO. __________ DATE ____________

IMPAIRMENT GRADING SCALE:

5: IMPERCEPTIBLE 4: PERCEPTIBLE BUT NOT ANNOYING
3: SLIGHTLY ANNOYING 2: ANNOYING 1: VERY ANNOYING

TEST SEQUENCE NO.	GRADE								
1	5		4		3		2		1
2	5		4		3		2		1
3	5		4		3		2		1
4	5		4		3		2		1
5	5		4		3		2		1
6	5		4		3		2		1
7	5		4		3		2		1
8	5		4		3		2		1
9	5		4		3		2		1
10	5		4		3		2		1
11	5		4		3		2		1
12	5		4		3		2		1
13	5		4		3		2		1
14	5		4		3		2		1
15	5		4		3		2		1
16	5		4		3		2		1
17	5		4		3		2		1
18	5		4		3		2		1

TEST SCORE SHEET

TABLE 6-1

	TEST NO.	EVALUATOR GROUP	TAPE NO.	CODEC TYPE	BIT RATE
DAY 1	1	A	2	L	128
	2	B	11	P	384
	3	A	10	P	256
	4	B	1	L	64
	5	A	6	H	768
	6	B	7	H	1536
LUNCH	---	---	---	---	---
	7	A	8	P	64
	8	B	3	L	256
	9	A	4	L	384
	10	B	9	P	128
	11	A	7	H	1536
	12	B	5	H	384
DAY 2					
	13	A	3	L	256
	14	B	8	P	64
	15	A	9	P	128
	16	B	4	L	384
	17	A	5	H	384
	18	B	6	H	768
	19	A	1	L	64
LUNCH	---	---	---	---	---
	20	B	2	L	128
	21	A	11	P	384
	22	B	10	P	256

Table 6 - 2 Test Schedule

6.4 Results

The results of all subjective tests are listed on Tables 6-3 to 6-13, one table for each codec and bit rate evaluated. The bottom right number is the mean overall rating. The individual scores were scrutinized to check them for overall consistency and to determine any erratic values. As expected, all scores vary over a wide range, depending on test sequence contents and evaluator. Each evaluator obviously had to form his/her own interpretation of the impairment grades which is typical for all subjective tests. Some evaluators tend to give lower grades than others but none could be identified as being consistently the lowest or highest scorer. The various test sequences were deliberately designed to provide different levels of stress on the codec algorithm, therefore the variation of scores between sequences confirms that the test tape is serving its intended purpose.

There were just 3 individual scores which seemed completely out of line as shown on Tables 6-3 and 6-7. Though they all were made by the same evaluator, the other scores of this individual were fully consistent, therefore there is no reason for invalidating all his scores. The scores in question probably are simply due to errors. Re-calculating the mean scores after eliminating the questionable numbers resulted in only a negligible change of the overall rating. All affected numbers are circled on the two tables and the changed values written in. This proves that there is sufficient variety in both test material and personnel that there can be high confidence in the validity of the results.

6.5 Discussion

A graphic summary of the mean scores of the three evaluated codecs is shown on Figure 6.3. The results are in full agreement with expectations. All scores improve with increasing bit rates. The shapes of the curves are somewhat different but do not deviate far from a straight line when using the conventional logarithmic bit rate scale. There is a noticeable difference between codecs L and P which operate over the same range of bit rates. It appears that codec L is optimized for performance at the high end of the bit rate range while codec P is better suited for use at the lower bit rates. The scores of codec H should not be directly compared with the others because they were obtained with a different and much more challenging test tape. The steepness of the curve indicates that codec H also was optimized for high bit rate performance. It was originally designed for operation at 1536 Kbps but its range was extended down as far as 384 Kbps.

L-64 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	2	2	2.5	1.5	1	1	1	2	2	1.67	.53
2	1	1	2	1	1	1	1	1	1	1.11	.31
3	2	2	3.5	2	2	2	2	3	2.5	2.33	.53
4	1	1	1.5	1	1	1	1	1	1	1.06	.16
5	1	1	1	1	1	1	1	2	1	1.11	.31
6	1	1	1.5	1	1	1	1	1	2	1.17	.33
7	1	2	2	1	1	1	1	2	1	1.33	.47
8	1	5	1	1	1	1	1	1	1	1.00 1.44	1.26 .00
9	1	1	1.5	1	1	1	2	1	1	1.17	.33
10	1	1	1.5	1	1	1	1	1	1	1.06	.16
11	1	1	1.5	1	1	1	1	1	1	1.06	.16
12	1	1	1.5	1	1	1	1	2	2	1.28	.42
13	1	1	2	1	2	1	1	2	3	1.56	.68
14	1	1	2	1	1	1	2	1	2	1.33	.47
15	1.5	2	3	1.5	3	1	4	3	3.5	2.50	.97
16	1.5	1	4	3	1	2	3	2	3.5	2.33	1.03
MEAN	1.19	1.27 1.50	2.00	1.25	1.25	1.13	1.50	1.63	1.78	1.47	
STD DEV	.35	1.00 .46	.83	.53	.56	.33	.87	.70	.90	1.44	

CODEC EVALUATION SUMMARY

CODEC L - 64 KBPS

TABLE 6-3

L-128 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	2	3	3	2	2	3	2	4	4.5	2.83	.88
2	1	2	2.5	2	1.5	2	2	2	2	1.89	.39
3	3	4	5	4	2.5	3	2	4	5	3.61	.99
4	2	2	2	1.5	2	1	1	1	1	1.50	.47
5	2	1	2.5	1	2	2	1.5	3	3	2.00	.71
6	2	2	2	2	2	1	1	1	1.5	1.61	.46
7	3	3	3	3	3	2	2	3	3.5	2.83	.47
8	1	1	2	1	1	1	1	1	1.5	1.17	.33
9	1	3	3.5	3.5	3.5	1	2	1	4.5	2.56	1.26
10	1	2	1.5	1	1.5	1	1	1	1.5	1.28	.34
11	1	1	2	2	1	1	1	3	3	1.67	.82
12	2	2	2.5	2	1.5	1	1	2	3	1.89	.61
13	3	3	2.5	3	3	2	1	3	4	2.72	.79
14	2	1	2.5	2	1	1	1	1	2	1.50	.58
15	2	3	3	4	4	2	3.5	3	4	3.17	.75
16	2	3	3	4	3	2	4.5	3	4.5	3.22	.89
MEAN	1.88	2.25	2.66	2.38	2.16	1.63	1.72	2.25	3.03	2.22	
STD DEV	.70	.90	.78	1.04	.90	.70	.98	1.09	1.27		

CODEC EVALUATION SUMMARY

CODEC L - 128 KBPS

TABLE 6-4

L-256 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	4	4	3.5	2.5	3	4	3	4	4	3.56	.55
2	2	2	3.5	2	2	3	2	3	3	2.50	.58
3	3	3	4	3.5	3	4	3	4	4	3.50	.47
4	3	2	2.5	1	2	2	1	3	3	2.17	.75
5	2	1	2.5	1.5	2	2	2	3	3	2.11	.61
6	3	1	2.5	1	2.5	2	1	2	4	2.11	.97
7	3	3	3.5	3	2	3	4.5	3	3	3.11	.61
8	3	1	2	1	2	2	1	2	2	1.78	.63
9	2	3	3.5	2.5	2	2	4	3	2.5	2.72	.67
10	2	1	3	2	2	2	2.5	2	2	2.06	.50
11	2	2	3	3	1.5	2	3.5	2	3	2.44	.64
12	2	2	3.5	2.5	1.5	2	3.5	3	3	2.56	.68
13	3	2	3.5	3	2	2	3	3	4	2.83	.67
14	3	1	3	2	2	1	3	2	3	2.22	.79
15	2.5	2	4	2	3	3	4	3	4	3.06	.76
16	3	3	4.5	3.5	3	4	4	3	4	3.56	.55
MEAN	2.66	2.06	3.25	2.25	2.22	2.5	2.81	2.81	3.22	2.64	
STD DEV	.58	.90	.64	.81	.50	.87	1.10	.63	.68		

CODEC EVALUATION SUMMARY

CODEC L - 256 KBPS

TABLE 6-5

L-384 CODEC EVALUATION SUMMARY

				EVALUATOR							STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	4	3	4	3.5	3	4	2	4	5	3.61	.81
2	3	4	3.5	3.5	2	4	4.5	4	4.5	3.67	.75
3	4	3	4.5	4	3	3	4	4	4.5	3.78	.58
4	3	2	3.5	3	2.5	3	3	4	4	3.11	.61
5	2	1	3	3.5	2	3	2	3	3	2.50	.75
6	2	2	3.5	3	2	4	2.5	3	3.5	2.83	.71
7	4	3	4	3	2	4	5	3	4	3.56	.83
8	3	1	2.5	2.5	3	3	2	2	3.5	2.50	.71
9	3	3	3.5	2.5	3	3	4	3	4.5	3.28	.58
10	3	3	3.5	3	3	3	3	3	3.5	3.11	.21
11	2	2	3.5	4	2	3	3.5	3	4.5	3.06	.86
12	3	2	4	3.5	2	2	3.5	2	4.5	2.94	.93
13	3	3	4	4	3	3	2.5	4	4.5	3.44	.64
14	3	2	3.5	3.5	3	2	2	3	4.5	2.94	.80
15	4	2	4	2.5	3	3	4	4	4.5	3.44	.80
16	4	2	4.5	4	3	3	4	4	5	3.72	.85
MEAN	3.13	2.38	3.69	3.31	2.59	3.13	3.22	3.31	4.22	3.22	
STD DEV	.70	.78	.50	.53	.47	.60	.95	.68	.56		

CODEC EVALUATION SUMMARY

CODEC L - 384 KBPS

TABLE 6-6

H-384 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	1	1	1.5	1	1	1	1	1	1	1.06	.16
2	1	5	1.5	1	1	1	1	1	1	1.06 1.50	1.25 .16
3	1	1	2	1	1	1	1.5	2	1	1.28	.42
4	1	1	1.5	1	1	1	1	1	1	1.06	.16
5	1	2	2	1	1	1	1	1	1	1.22	.42
6	1	1	1.5	1	1	1	1	1	1	1.06	.16
7	1	1	1.5	1	1	1	2.5	2	2	1.44	.55
8	1	1	1.5	1	1	1	1	1	1	1.06	.16
9	2	2	3	2	2	2	2.5	2	4	2.39	.66
10	1	1	2	1	1	1	1	1	1	1.11	.31
11	1	1	2	1	1	1	1	1	2	1.22	.42
12	2	1	2.5	2	2	1	1.5	1	3	1.78	.67
13	1	1	1.5	1	1	1	1	1	2	1.17	.33
14	1	1	1	1	1	1	2.5	2	1	1.28	.53
15	1	1	2	1	1	1	1	1	2	1.22	.42
16	1	1	2.5	1	1	1	1	1	3	1.39	.74
17	1	1	2	1	1	1	1.5	1	1	1.17	.33
18	1	5	1	1	1	1	1	1	1	1.00 1.44	1.26 .00
MEAN	1.13	1.14 1.38	1.84	1.13	1.13	1.06	1.34	1.25	1.69	1.33 1.30	
STD DEV	.33	.99 .36	.49	.33	.33	.24	.58	.43	.92		

CODEC EVALUATION SUMMARY

CODEC H - 384 KBPS

TABLE 6-7

H-384 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	1	1	1.5	1	1	1	1	1	1	1.06	.16
2	1	5	1.5	1	1	1	1	1	1	1.06 1.50	1.25 .16
3	1	1	2	1	1	1	1.5	2	1	1.28	.42
4	1	1	1.5	1	1	1	1	1	1	1.06	.16
5	1	2	2	1	1	1	1	1	1	1.22	.42
6	1	1	1.5	1	1	1	1	1	1	1.06	.16
7	1	1	1.5	1	1	1	2.5	2	2	1.44	.55
8	1	1	1.5	1	1	1	1	1	1	1.06	.16
9	2	2	3	2	2	2	2.5	2	4	2.39	.66
10	1	1	2	1	1	1	1	1	1	1.11	.31
11	1	1	2	1	1	1	1	1	2	1.22	.42
12	2	1	2.5	2	2	1	1.5	1	3	1.78	.67
13	1	1	1.5	1	1	1	1	1	2	1.17	.33
14	1	1	1	1	1	1	2.5	2	1	1.28	.53
15	1	1	2	1	1	1	1	1	2	1.22	.42
16	1	1	2.5	1	1	1	1	1	3	1.39	.74
17	1	1	2	1	1	1	1.5	1	1	1.17	.33
18	1	5	1	1	1	1	1	1	1	1.00 1.44	1.26 .00
MEAN	1.13	1.14 1.38	1.84	1.13	1.13	1.06	1.34	1.25	1.69	1.33 1.30	
STD DEV	.33	.99 .36	.49	.33	.33	.24	.58	.43	.92		

CODEC EVALUATION SUMMARY

CODEC H - 768 KBPS

TABLE 6-8

H-1536 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	4	3	4	3.5	2	3	3	4	4	3.39	.66
2	3	2	4	3	2	2	2	3	3	2.67	.67
3	3.5	3	4	4	4	4	5	4	4	3.94	.50
4	4	2	4	3	3	3	4	3	2	3.11	.74
5	4	2	4	4.5	3	3	5	4	4	3.72	.85
6	4	3	4	4	4	4	4	4	4	3.89	.31
7	3	1	3	2	2	3	4.5	4	3	2.83	1.00
8	4	3	3	3.5	4	3	3.5	4	3	3.44	.44
9	4	3	4.5	4	4	3	5	4	4	3.94	.60
10	3.5	3	3	2	3	3	3	3	2	2.83	.47
11	4	4	4	3	3.5	4	4	3	4	3.72	.42
12	4	2	3.5	3.5	3	3	4.5	4	4	3.50	.71
13	4	4	4	3.5	4	4	4	4	4	3.94	.16
14	3.5	3	4	4	3.5	4	5	4	4	3.89	.52
15	4	4	4.5	4.5	3	4	5	4	4	4.11	.52
16	4	3	3.5	4	4	4	4.5	3	4	3.78	.48
17	3	2	3	3	2	3	4	3	4	3.00	.67
18	4	3	4	4	3	3	4.5	3	3	3.50	.58
MEAN	3.78	2.81	3.81	3.50	3.25	3.38	4.13	3.69	3.56	3.55	
STD DEV	.35	.81	.46	.73	.73	.60	.86	.46	.70		

CODEC EVALUATION SUMMARY

CODEC H - 1536 KBPS

TABLE 6-9

P-64 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	4	3	4	4	3	3	3	4	4	3.56	.50
2	1	2	3	1	1	2	2	2	3	1.89	.74
3	4	2	4.5	4	3	2	2	4	4	3.28	.97
4	1	1	1.5	1	1	1	1	1	1.5	1.11	.21
5	1	1	1.5	1	1	1	1.5	2	1.5	1.28	.34
6	1	1	2	1	1	1	1	1	3	1.33	.67
7	2	2	2.5	3	2	2	1	3	3	2.28	.63
8	1	1	1.5	1	1	1	1	1	1.5	1.11	.21
9	2	1	2.5	1	1.5	2	1	3	3	1.89	.77
10	1	1	1.5	1	1	1	1	2	2	1.28	.42
11	1	2	1.5	1	1	2	2	2	2	1.61	.46
12	1	1	2.5	1	2	1	1	2	3	1.61	.74
13	5	2	2.5	3	2	2	1	3	4	2.72	1.13
14	2	1	1.5	1	1	1	1	1	3	1.39	.66
15	3	2	3.5	2	3	3	4	4	4.5	3.22	.82
16	3	3	4.5	3	2	3	2	4	5	3.28	.97
MEAN	2.06	1.63	2.53	1.81	1.66	1.75	1.59	2.44	3	2.05	
STD DEV	1.30	.70	1.05	1.13	.76	.75	.85	1.12	1.06		

CODEC EVALUATION SUMMARY

CODEC P - 64 KBPS

TABLE 6-10

P-128 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	4	4	4	4	3	4	3.5	3	3	3.61	.46
2	3	2	3.5	3	3	2	3	3	3	2.83	.47
3	4	3	4.5	4.5	4	2	3.5	3	4	3.61	.77
4	1	1	3	2	1	1	1	2	3	1.67	.82
5	1	1	3	2	1	1	3	2	3	1.89	.87
6	2	1	3	2	2	2	2	2	3	2.11	.57
7	2.5	2	3.5	2.5	3	3	4	3	3	2.94	.55
8	1	1	3	1.5	2	1	1	2	1	1.50	.67
9	3	2	4	1.5	2	2	3.5	3	2	2.56	.80
10	2	2	3	1.5	2	2	3.5	3	3	2.44	.64
11	1	1	3	1.5	1	2	3	2	2	1.83	.75
12	2	1	3.5	2	1	2	2	2	2	1.94	.68
13	3	2	3.5	3	2	3	5	3	4	3.17	.88
14	1	1	2.5	1.5	1	2	1	2	2	1.56	.55
15	3	3	4	3	3	3	4.5	4	4	3.50	.58
16	3	2	4	4	3	2	4	3	4	3.22	.79
MEAN	2.28	1.81	3.44	2.47	2.13	2.13	2.97	2.63	2.88	2.52	
STD DEV	1.03	.88	.53	.98	.93	.78	1.21	.60	.86		

CODEC EVALUATION SUMMARY

CODEC P - 128 KBPS

TABLE 6-11

P-256 CODEC EVALUATION SUMMARY

	EVALUATOR										STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	4	3	4	4	4	3	3.5	4	4	3.72	.42
2	2	3	3.5	3	3	2	2.5	3	4.5	2.94	.72
3	4	3	4.5	4	3	3	3.5	4	4.5	3.72	.58
4	4	3	3.5	4	2.5	2	3.5	4	4	3.39	.70
5	2	2	3.5	3	3	2	3	3	4	2.83	.67
6	3	1	3.5	2.5	2.5	2	3	3	3.5	2.67	.75
7	4	2	4	3.5	2.5	2	3.5	4	3.5	3.22	.79
8	2	1	3	2.5	2.5	2	2	3	3	2.33	.62
9	3	3	3	3.5	3	2	3.5	3	2.5	2.94	.44
10	2	2	3	3	3	2	2	3	2	2.44	.50
11	2	2	3	2	2	3	3.5	3	3.5	2.67	.62
12	3	2	3	3	2	2	3	3	3.5	2.72	.53
13	4	3	4	4	3.5	3	4	3	3.5	3.56	.44
14	3	2	2.5	2	2.5	2	2.5	2	2.5	2.33	.33
15	4	2	4.5	3	3.5	3	3	4	4.5	3.50	.78
16	4	2	4.5	4	3	3	4	3	4	3.50	.75
MEAN	3.13	2.25	3.56	3.19	2.84	2.38	3.13	3.25	3.56	3.03	
STD DEV	.86	.66	.61	.68	.52	.48	.60	.56	.73		

CODEC EVALUATION SUMMARY

CODEC P - 256 KBPS

TABLE 6-12

P-384 CODEC EVALUATION SUMMARY

				EVALUATOR							STD
SEQ	1	2	3	4	5	6	7	8	9	MEAN	DEV
1	4	3	4.5	4	3	3	1	4	4	3.39	.99
2	3	3	4	3	2	2	3	4	4	3.11	.74
3	4	3	4.5	4	3	2	4	4	4	3.61	.74
4	3	2	3.5	2	3	3	4	3	3	2.94	.60
5	3	2	4	2	2	3	4	3	3	2.89	.74
6	3	2	3.5	4	2.5	2	4	4	4	3.22	.82
7	3.5	2	4	4	3	2	2	3	4	3.06	.83
8	3	3	3	3	2.5	2	2	3	3	2.72	.42
9	3	2	3.5	4	3	2	4	3	4	3.17	.75
10	3	3	4	4	3	3	3	3	3	3.22	.42
11	3	1	3.5	4	2.5	3	2	2	2	2.56	.86
12	4	2	4	4	3	3	3	2	3	3.11	.74
13	4	2	4.5	4.5	3.5	4	4	3	5	3.83	.85
14	3.5	2	3.5	3	3	2	3	2	4	2.89	.70
15	4	2	4	4	4	3	5	4	4	3.78	.79
16	4	2	4.5	3.5	2	3	4	3	4	3.33	.85
MEAN	3.44	2.25	3.91	3.56	2.81	2.63	3.25	3.13	3.63	3.18	
STD DEV	.46	.56	.44	.73	.53	.60	1.03	.70	.70		

CODEC EVALUATION SUMMARY

CODEC P - 384 KBPS

TABLE 6-13

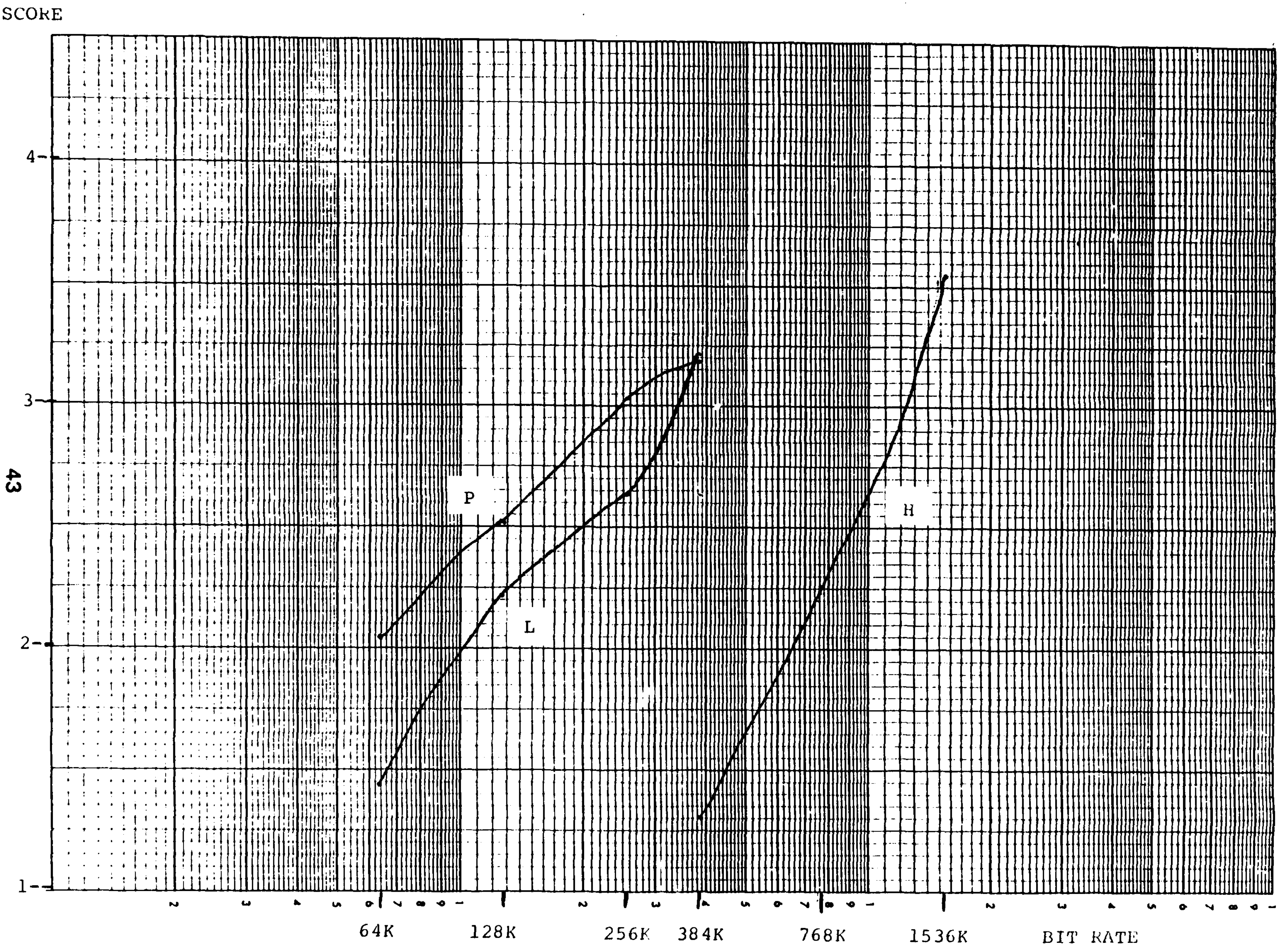


FIGURE 6.3 CODEC RATING SUMMARY

It must be emphasized that the subjective score reflects the overall impression of the test picture on each evaluator. The most common impairments can be roughly put into 3 categories, smearing, blocking and jerkiness. Each evaluator is likely to unknowingly put different levels of emphasis on each of these categories which is partly responsible for the differences in scores. Not all impairments exist in every codec. Only the algorithm of codec H produces blocking which is mainly prevalent at low bit rates which explains the unusually low score at 384 Kbps. Codec H operates at a fixed transmitted frame rate which produces a small and constant amount of jerkiness independent of bit rate. Codecs L and P operate at variable transmitted frame rates, depending on bit rate and the amount of detail and motion in the test picture. Thus, the resulting jerkiness becomes an important factor in picture quality assessment. Codec manufacturers generally make trade-offs between smearing and jerkiness to achieve what they consider the best overall results.

The reader will recall that the distortion terms associated with scores 2 and 3 are "annoying" and "slightly annoying" respectively. It may be therefore concluded that the threshold of acceptability for teleconferencing purposes may occur at a score of approximately 2.5. The results shown in Figure 6.3 appear reasonable since low bit rate codecs have been accepted in the market place at approximately 128 Kbps and high bit rate codecs have been accepted at 768 Kbps and higher. This is essentially identical with the test results presented in Section 5.0. It should also be noted that CCIR Rec. 500-3 suggests a grand mean score of 3 which logically should represent a picture quality well above the threshold of acceptability.

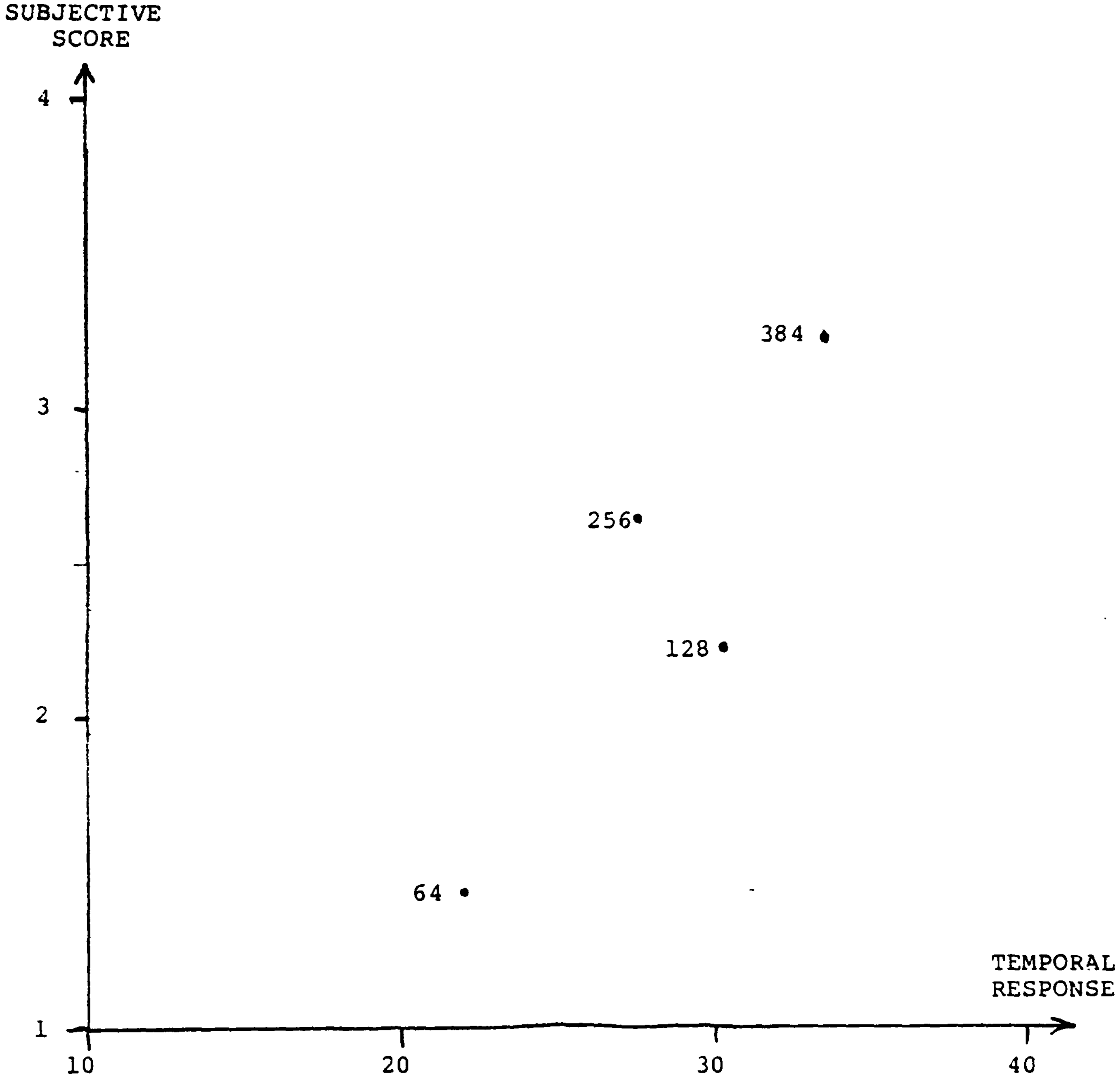
6.6 Correlation with Objective Tests

Objective tests of teleconferencing codecs fall into two categories, still picture and motion picture tests. Still picture tests are conventional and easily implemented while objective motion tests are still in an early stage of development. Previous experiments have shown that subjective rating of teleconferencing picture is primarily dependent on motion rendition with most still picture parameters having much less influence. Therefore, the tests described in this report have concentrated on motion performance. The present status of objective motion testing is not sufficiently advanced to expect perfect correlation but whatever results can be obtained will be significant in providing guidance for further development of objective motion testing techniques.

Conventional static video tests give firm results independent of the test signals and methods that were used. This is not the case with objective motion tests. The methodologies developed so far using the rotating wheel pattern give numerical values of temporal response and transmitted frame rate for each codec and bit rate but these results vary considerably with the spoke width and rotation speed of the pattern. There is no obvious reason for selection of any particular pattern to be optimal for correlation with subjective results. A choice had to be made based on availability of temporal response and transmitted frame rate data over the whole range of bit rates requiring a minimum of averaging, interpolation and/or extrapolation. This somewhat arbitrary choice is the 18° spoke width and the temporal response value at a temporal frequency of 2 cycles per second. These values are close to the center of the full range of the objective tests which is the same as for the subjective tests described herein.

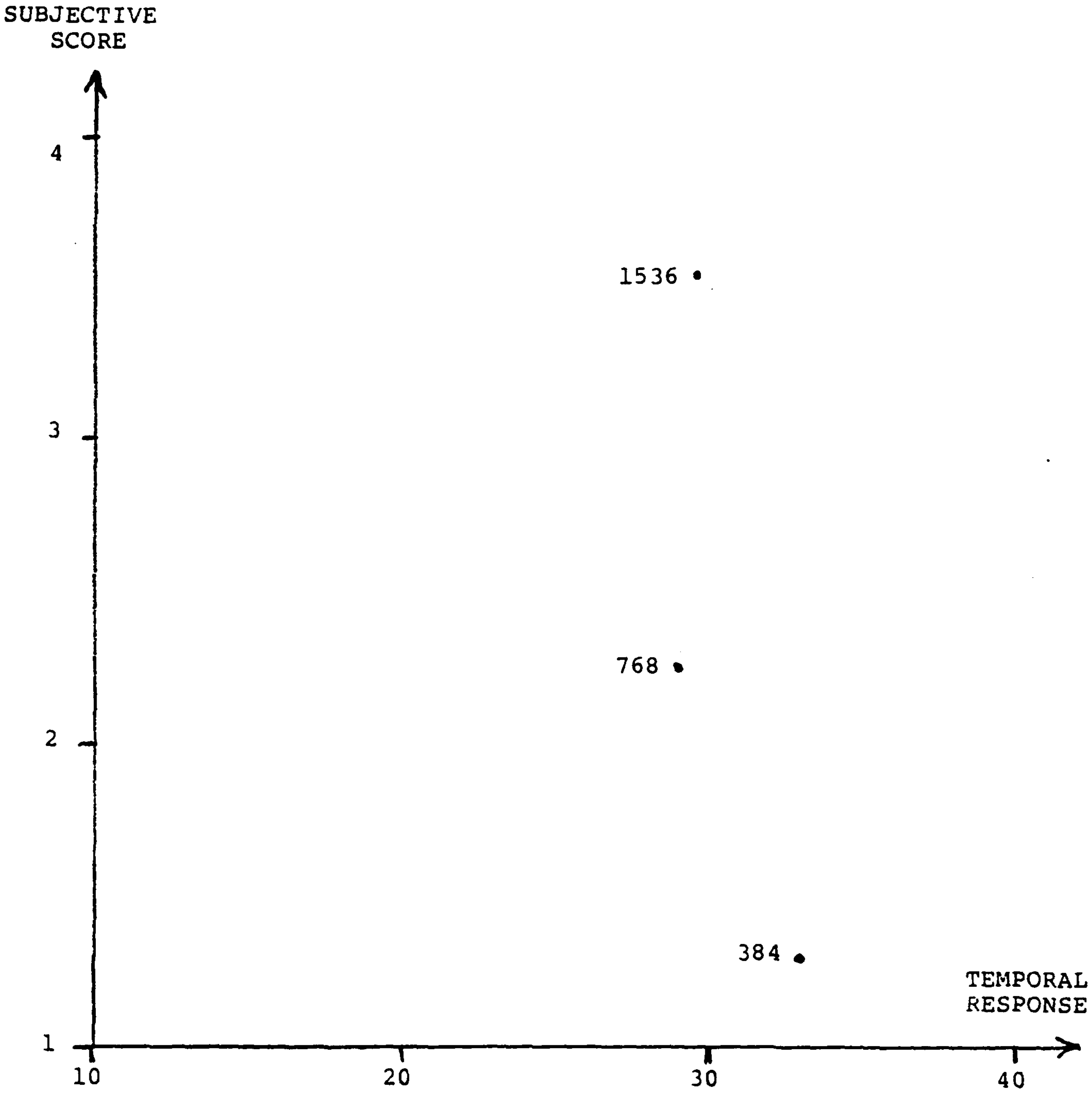
The resulting correlation points are shown on Figures 6.4 to 6.6 for temporal response and Figures 6.7 and 6.8 for transmitted frame rate which is a measure of jerkiness. Each figure shows the results for one codec model. Since Codec H has a fixed transmitted frame rate of 15, correlation with this parameter would be meaningless. The transmitted frame rate values for the other two codecs are the rounded off averages over the measurement range. A correlation point is shown for each bit rate. No attempt has been made to draw a correlation curve since at present there is no basis to establish a theoretical line of 100% correlation.

In general, the results are not ideal but reasonable and useful. The correlation with the transmitted frame rate values is mostly good, showing a consistent increase of both parameters with bit rate. The correlation with the temporal response indicates the same tendency but each low rate codec shows a reversal at one bit rate. This is most likely due to imperfections in the method of objective measurement of temporal response which is still in need of further refinement. The results with the high rate codec are less consistent, indicating only little change of temporal response with bit rate. This could partially be due to the choice of test pattern and sample point for temporal response but is more likely to be caused by the occurrence of heavy blocking, particularly at 384 Kbps, which is not well enough recognized by the present objective test method. Thus, one important result of these tests is that the presently used technique to measure temporal response should be refined and expanded to include improved recognition



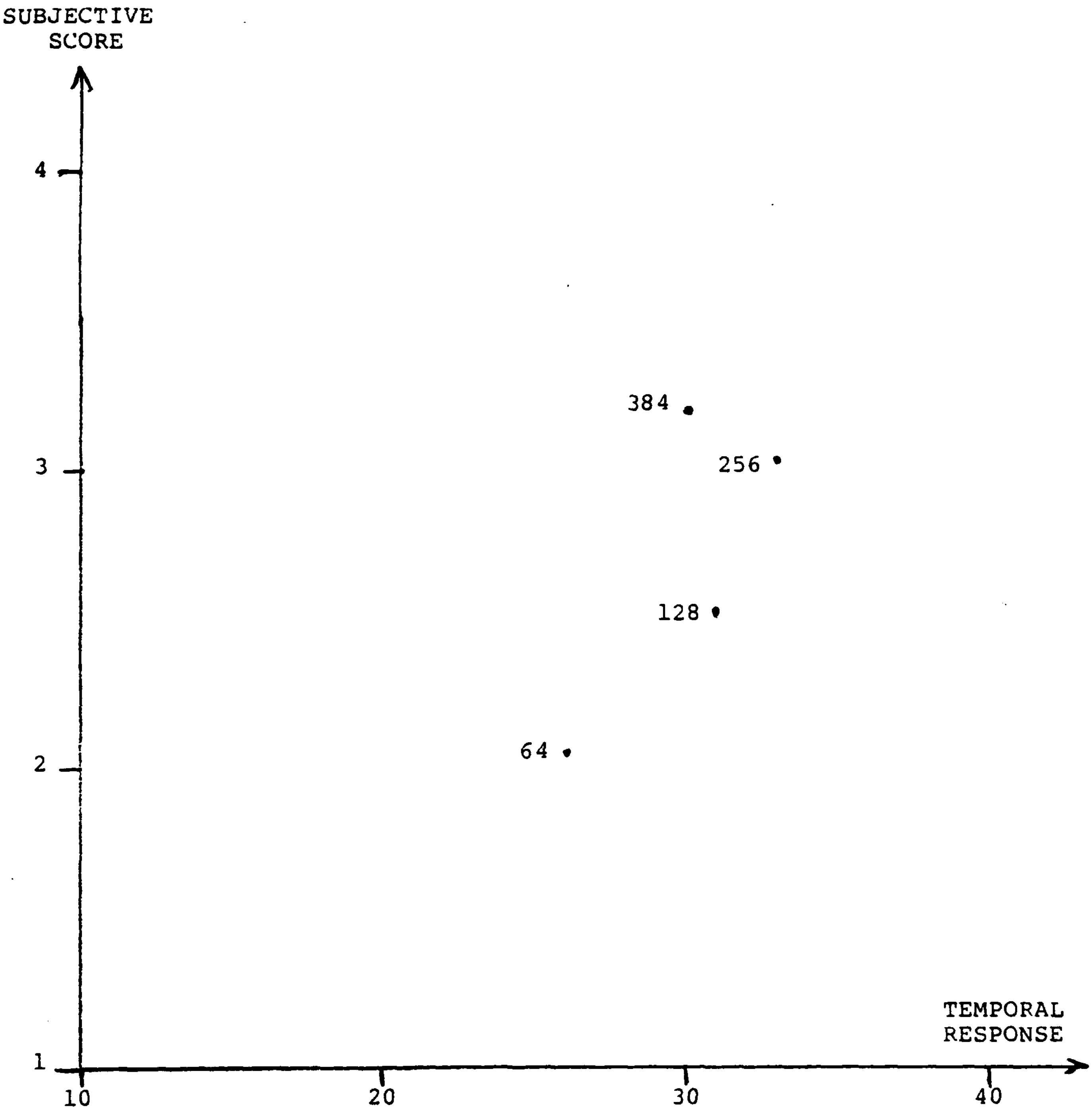
CORRELATION - SUBJECTIVE SCORE VS. TEMPORAL RESPONSE - CODEC L

FIGURE 6.4



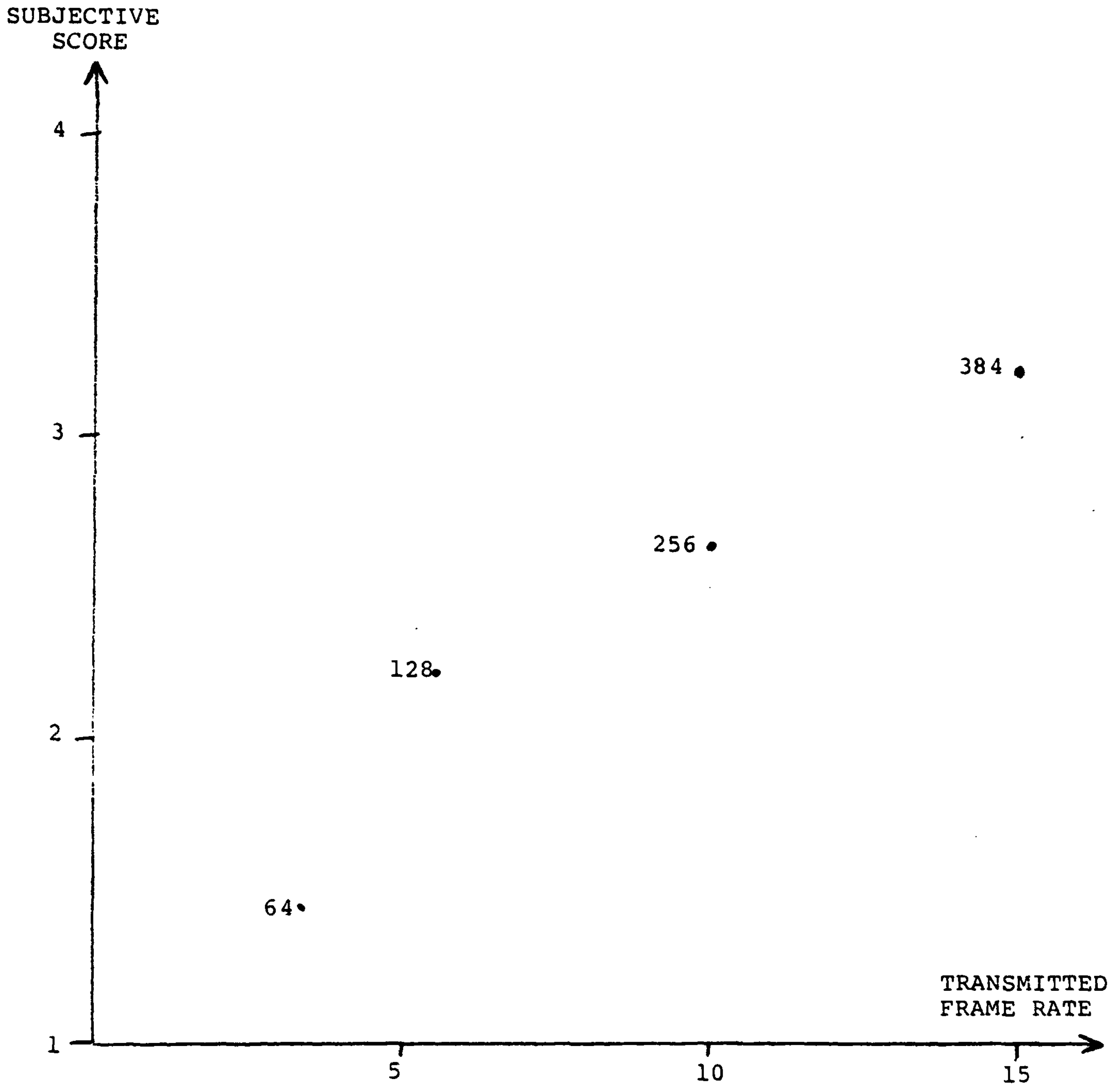
CORRELATION - SUBJECTIVE SCORE VS. TEMPORAL RESPONSE - CODEC H

FIGURE 6.5



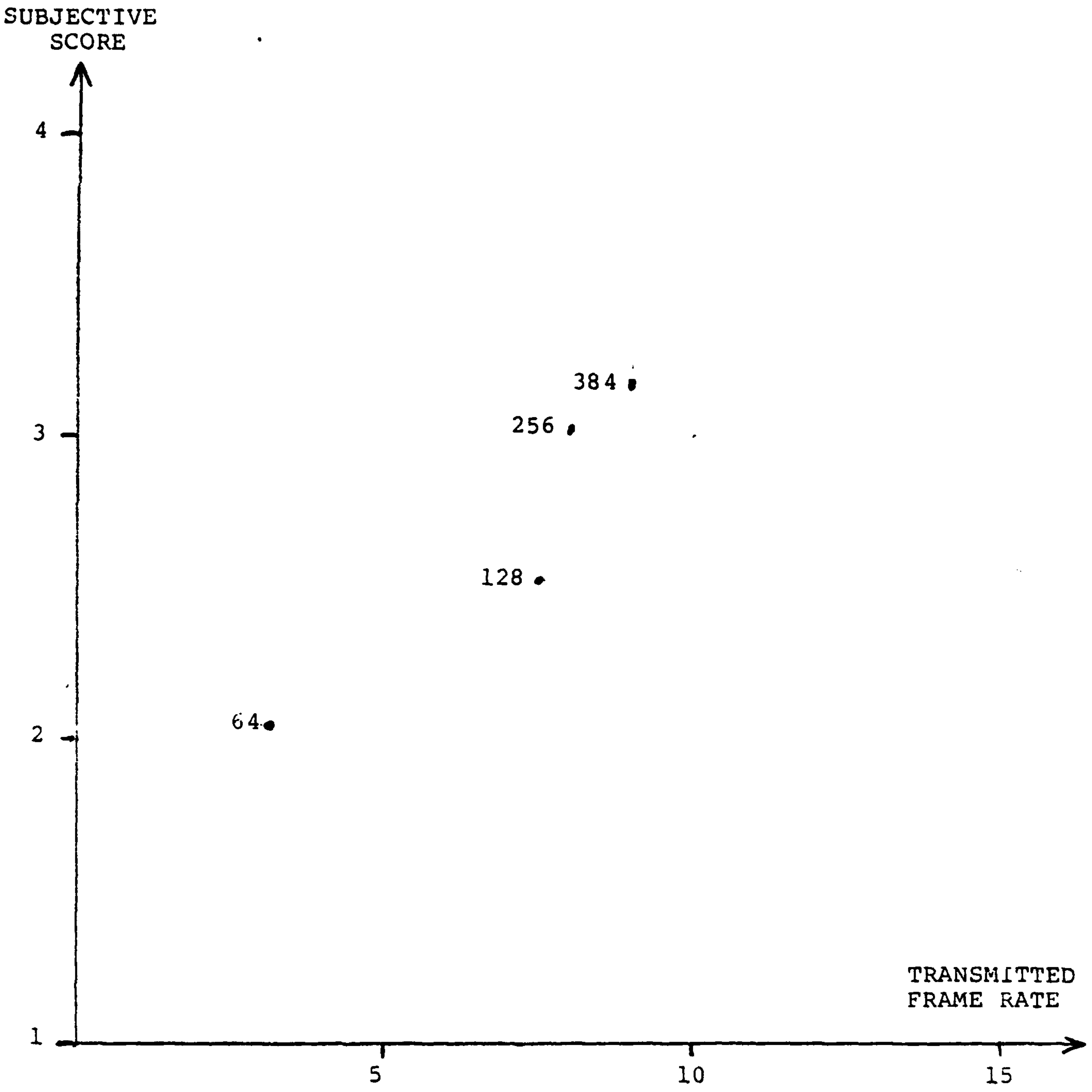
CORRELATION - SUBJECTIVE SCORE VS. TEMPORAL RESPONSE - CODEC P

FIGURE 6.6



CORRELATION - SUBJECTIVE SCORE VS. TRANSMITTED FRAME RATE (JERKINESS) - CODEC L

FIGURE 6.7



CORRELATION - SUBJECTIVE SCORE VS. TRANSMITTED FRAME RATE (JERKINESS) - CODEC P

FIGURE 6.8

and measurement of blocking and other individually definable and measurable artifacts.

7.0 CONCLUSION AND RECOMMENDATIONS

The effort described in this report covers the methodology and results of testing digital video teleconferencing codecs by means of specially developed test tapes processed through a codec and then submitted to a group of evaluators for quality rating. The results are tabulated and analyzed and found to be logical and consistent. The test tapes that were used showed scenes with various amounts of motion, therefore, the results mainly depict codec motion performance. Correlation with the results of objective test methods for moving pictures presently under development shows acceptable results and promise for the future considering that this is a first attempt in this still unexplored field.

The test material used in this program emphasizes scenes with "live" motion content. Further sections of processed test tapes are available featuring still graphics and graphics with motion such as marking and pointing. These tapes should also be evaluated subjectively using the same method as for this program, and the test scores correlated with the available results of conventional still picture tests. This will show which of the many still picture test parameters are relevant for the objective evaluation of codec pictures.

A subsequent longer term program should concentrate on the improvement of objective motion test methods. This program has shown that present tests do not sufficiently recognize blocking which is an important artifact in many codec algorithms, particularly units built to the new CCITT H.261 Recommendation. Present objective test methods must be further developed to produce better correlation of subjective and objective test results for all codecs at all operating bit rates.

The subjective general picture quality evaluations which represent the main output of this program are of fundamental importance but by no means the only criteria to be considered. It is of equal importance to sort out the various test sequences and assign each to one or several user application groups still to be agreed upon. Different levels of quality will be needed depending on the specific application. It will subsequently be possible to establish both subjective and objective thresholds of acceptability for each application group.

Finally, new codecs in accordance with CCITT H.261 Recommendation are becoming available both in the USA and overseas. Pictures on some of these units have been viewed but so far no formal independent tests have been performed. It is important that such tests be implemented as soon as several designs of these equipments become available.